Mean-Field Microcanonical Gradient Descent

Marcus Häggbom
SEB Group *and*
KTH Royal Institute of Technology
Stockholm, Sweden

Morten Karlsmark
SEB Group
Stockholm, Sweden

Joakim Andén
KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

Microcanonical gradient descent is a sampling procedure for energy-based models allowing for efficient sampling of distributions in high dimension. It works by transporting samples from a high-entropy distribution, such as Gaussian white noise, to a low-energy region using gradient descent. We put this model in the framework of normalizing flows, showing how it can often overfit by losing an unnecessary amount of entropy in the descent. As a remedy, we propose a mean-field microcanonical gradient descent that samples several weakly coupled data points simultaneously, allowing for better control of the entropy loss while paying little in terms of likelihood fit. We study these models in the context of stationary time series and 2D textures.

1 INTRODUCTION

The defining characteristic of a well-behaved generative model is the balance between its ability to, on the one hand, produce samples that are typical of the training data, while on the other hand having a significant amount of diversity within its samples. For example, a generative adversarial network (GAN) which has suffered mode collapse could produce great samples within one mode but not others. Similarly, the empirical distribution of the training data approximates the training data well but is useless for generating new samples, while a Gaussian white noise model may produce highly diverse samples that have no relation to the training data. Formally, we can view this in terms of the reverse Kullback–Leibler (KL) divergence (Papamakarios et al., 2021) of the generative model q with respect to the true distribution p on the sample space $\mathcal{X}$:

$$\mathcal{D}_{\mathrm{KL}}(q \parallel p) = -H(q) - \mathbb{E}_q[\log p(X)], \tag{1}$$

where $H(q)$ denotes the differential entropy of q and $\mathbb{E}_q$ is the expected value with respect to q. To achieve a good fit, that is, a low KL divergence, we thus want to simultaneously maximize the entropy $H(q)$ and the log-likelihood $\mathbb{E}_q[\log p(X)]$ of p under the approximation q.

One popular family of generative models is that of the energy-based model (EBM) (Geman and Geman, 1984), also known as a *canonical* or *macrocanonical ensemble* (Jaynes, 1957), typically formulated as the Gibbs or Boltzmann distribution $q(x) \propto \exp(-\beta \cdot \Phi(x))$ for a energy function $\Phi : \mathcal{X} \to \mathbb{R}^K$ and parameter vector $\beta \in \mathbb{R}^K$ (the inverse temperature). This is the distribution that maximizes the entropy $H(q)$ subject to the moment constraint $\mathbb{E}_q[\Phi(X)] = \alpha$ for some target energy vector $\alpha \in \mathbb{R}^K$ (Cover and Thomas, 2006).

In this work, we tackle the one-shot learning problem, where we are given Φ and $\alpha = \Phi(y)$ is obtained from some observation $y \in \mathcal{X}$. Here, Φ may be given by some domain-specific design or earlier learning procedures. Using the macrocanonical approach here suffers from two main challenges, namely determining β and sampling, both nontrivial in the general case and in particular when $\mathcal{X}$ is high-dimensional. As a remedy to the first issue is the *microcanonical ensemble* (Lanford, 1975; Ellis et al., 2000; Touchette, 2015), which is also a maximum-entropy distribution but constrained to distributions with support in the *microcanonical set* of width $\varepsilon > 0$,

$$\Omega_\varepsilon \coloneqq \{x \in \mathcal{X} : \|\Phi(x) - \alpha\| \leq \varepsilon\}. \tag{2}$$

Maximizing the entropy here implies that the distribution is uniform over this set. Thus, the entropy is equal to the log of the volume of Ω_ε which is increasing in ε. This approximation relies on the assumption that $\Phi(X)$ concentrates around its mean with high probability under the true distribution p, which is the case for

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

most stationary time series of sufficiently long duration and when Φ is defined as the time average of time-shift equivariant potentials. The parameter ε can then be adjusted to match this concentration of $\Phi(X)$.

While the microcanonical ensemble avoids the issue of estimating β in the macrocanonical model, sampling in high-dimensional spaces remains challenging. To mitigate this, the microcanonical gradient descent model (MGDM) was introduced by Bruna and Mallat (2019) as an approximation of the microcanonical ensemble which is easier to sample from, and has been successfully applied in a variety of domains (Bruna and Mallat, 2019; Leonarduzzi et al., 2019; Morel et al., 2023; Zhang and Mallat, 2021; Brochard et al., 2022b; Cheng et al., 2024; Auclair et al., 2023). The MGDM is defined as the pushforward of Gaussian white noise by way of a sequence of gradient descent steps that seek to minimize the objective

$$L(x) \coloneqq \frac{1}{2}\|\Phi(x) - \alpha\|^2. \tag{3}$$

Thus, taking $\mathcal{X} = \mathbb{R}^d$, samples from the MGDM are generated by sampling x_0 from $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ for some initial variance σ^2 and updating the sample using

$$x_{t+1} = g(x_t) \coloneqq x_t - \gamma J_\Phi^\top(x_t)(\Phi(x_t) - \alpha), \tag{4}$$

where γ is the step size and $J_\Phi(x) \in \mathbb{R}^{K \times d}$ is the Jacobian of Φ in x. This is typically iterated for a fixed number of steps T or until x_t reaches the microcanonical set Ω_ε for some fixed ε (Leonarduzzi et al., 2019; Morel et al., 2023).

Despite its success, MGDM can be shown to suffer from entropy collapse in many cases, resulting in a model that is able to produce typical samples but lacks sufficient variability. We shall see that this is due to the contraction of the distribution that typically occurs with each gradient step, reducing the entropy and leading to a higher KL divergence. To remedy this, we propose a new variant of the MGDM, called the mean-field microcanonical gradient descent model (MF–MGDM), which generates a batch of samples $\boldsymbol{x} \coloneqq \{x^{(n)}\}_{n=1}^N$ such that their mean energy vector satisfies the necessary constraints, effectively replacing Φ in (3) with the batch mean

$$\overline{\Phi}(\boldsymbol{x}) \coloneqq \frac{1}{N}\sum\nolimits_{n=1}^{N} \Phi(x^{(n)}). \tag{5}$$

In this model, the initial distribution is not so much contracted as transported through the energy space to the target while maintaining more of its initial entropy. We provide a theoretical justification for this in the form of a tighter lower bound on the entropy. The resulting model combines the expressiveness of the micro- and macrocanonical ensembles with the efficient sampling of the MGDM. The choice of energy function Φ is highly dependent on the particular distribution to be approximated. To illustrate the power of the proposed approach, we therefore evaluate MF–MGDM for a range of possible functions. In each case, we see a significant improvement over the basic MGDM approach, validating the theoretical results obtained on the entropy lower bound.

The structure of this article is as follows. Section 2 surveys the literature on energy-based models and the MGDM in particular, while Section 3 illustrates the entropy collapse observed in the MGDM. A proposed solution to this is introduced in Section 4 in the form of the MF–MGDM along with a lower bound on its entropy, and numerical results supporting this algorithm are presented in Section 5. Python code to reproduce the results in this paper may be found at `https://github.com/MarcusHaggbom/mf-mgdm`.

2 RELATED WORK

The micro- and macrocanonical ensembles are both maximum entropy distributions conditioned on the target energy α. These are related via the Boltzmann equivalence principle (Lanford, 1975), which states that under certain conditions of Φ, they converge to the same measure as $\dim \mathcal{X} \to \infty$ and $\varepsilon \to 0$. While it is not guaranteed that a maximum entropy distribution exists in the macrocanonical case (Cover and Thomas, 2006), the microcanonical ensemble is more general in that it allows for a wider range of energy functions (Strominger, 1983; Bruna and Mallat, 2019). Both ensembles allow sampling by MCMC methods, which is computationally challenging, but have been employed in high dimensions for EBMs (Du and Mordatch, 2019) and score-based diffusion models (Yang et al., 2023). This relies on sufficient mixing of the Markov chain, which is crucial for obtaining reliable Monte Carlo estimates in finite time, e.g. of the expectations in (1) when comparing models with respect to the reverse KL divergence.

The MGDM was introduced in Bruna and Mallat (2019) for the purpose of facilitating sampling. Each step is deterministic, allowing us to calculate the exact likelihood of each sample, which, unlike MCMC methods, makes computing entropy comparatively easy. The MGDM has been used in a variety of applications, such as cosmology (Eickenberg et al., 2022; Cheng et al., 2024; Auclair et al., 2023) and texture synthesis (Brochard et al., 2022b; Zhang and Mallat, 2021). In these contexts, the model is often paired with various extensions of the *scattering transform* (Mallat, 2012) used as features in the energy function. The scattering transform is a composition of wavelet transforms and non-linearities, and can be seen as a convolutional neural net with

predefined weights (Mallat, 2016). Apart from its use as an energy function in generative models, it has also found applications in image classification (Bruna and Mallat, 2013; Villoutreix et al., 2017; Oyallon et al., 2019), audio similarity measurement (Lostanlen et al., 2018; Andén et al., 2019; Lostanlen et al., 2021), molecular energy regression (Eickenberg et al., 2017), and heart beat classification (Chudáček et al., 2013a,b, 2014; Warrick et al., 2020) among others. Appendix D.1 introduces the scattering transform in more detail.

Independently of the development of this work, the idea of optimizing over multiple samples to increase energy variance in MGDM had been used previously by Allys et al. (2020) and Cheng et al. (2024) to synthesize physics data. The former draws similar conclusion as we do on improved variability in energy – see (Allys et al., 2020, Fig. 9) where marginal empirical energy distributions are plotted – whereas the latter states that the approach makes the model "closer to its macrocanonical limit" (Cheng et al., 2024). However, neither work performed a theoretical analysis to justify and evaluate this claim.

MF–MGDM can be seen as an instance of the general particle gradient descent method, which is studied by lifting the optimization to the space of measures, and where the particle updates are seen as an update of the corresponding empirical measure. It is analyzed by Chizat and Bach (2018) in the continuous and mean-field limit settings, and proposed in a stochastic version by Nitanda and Suzuki (2017) for infinite ensembles of neural network classifiers. A related method is the Stein variational gradient descent by Liu and Wang (2016). Here, an empirical measure is transported by iteratively minimizing the KL divergence over a certain function space of perturbations, where the optimal update is an expectation with respect to the current measure.

Another instance of particle gradient descent is due to Brochard et al. (2022a), who extends the MGDM for the purpose of modeling point processes. The discretized versions of these processes result in a set of binary random variables in a lattice, which means the MGDM cannot be applied as is since the gradient is not defined. The same situation arises for the Ising model (Lenz, 1920), which Bruna and Mallat (2019) circumvents by relaxing the optimization over real values but imposing the binary constraint by adding a penalty term to the objective (3). Brochard et al. (2022a) instead leverages the idea of representing the point processes as random measures taking the form of sums of point masses, whose locations are updated continuously with particle gradient descent. The resulting update scheme bears resemblance with MF–MGDM, but is fundamentally different since it is defined on a space of (counting) measures. For instance, the microcanonical set is in this case not defined as a set of points, but a set of measures. Going one step further, this also means that it is possible to define a mean-field version of their model.

In the context of finance, MGDMs coupled with variants of the scattering transform have been used to generate sample paths of time series. In Leonarduzzi et al. (2019), it is shown that the time-average of the second-order scattering transform encodes heavy tails, and that including also phase harmonic correlations (Mallat et al., 2019) encapsulates temporal asymmetries, both of which are typical features of financial time series. An extension of this representation is the scattering spectrum (Morel et al., 2023), which increases sparsity and better captures multiscale properties of rough paths such as fractional Brownian motion.

Another popular feature representation for rough paths is the truncated *signature* (Lyons, 2014). The full signature of a path is a lossless representation up to time parametrization, and the truncation error decreases as the inverse of the factorial of the number of included terms. Whereas the features based on the scattering transform are typically used as is, the truncated signature usually functions as a compact initial feature on top of which learning methods are applied. In financial time series generation, this encoding has proved efficient for other generative models, e.g. variational autoencoders (Buehler et al., 2020) and Wasserstein GANs (Ni et al., 2021). In principle, these learned features could serve as energy function in the canonical ensembles.

3 OVERFITTING TO TARGET ENERGY

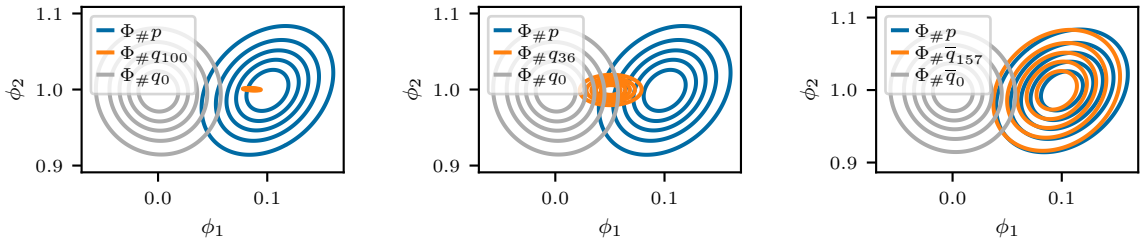
With each gradient step, the MGDM pushes the energy vector $\Phi(x)$ of a sample x from the initial distribution towards the target energy α. In doing so, however, the distribution of x and $\Phi(x)$ also contracts. As a result, by the time the process reaches the microcanonical set Ω_ε, a significant reduction of entropy has been incurred, producing a poor fit to the microcanonical ensemble.

3.1 An Illustrative Example

As an example, we consider the AR(1) model with parameter φ and conditional variance σ^2:

$$x_i = \varphi x_{i-1} + \sigma\varepsilon_i, \tag{6}$$

where $(\varepsilon_i)_i$ is Gaussian white noise. If $|\varphi| < 1$, the process is stationary and has the marginal distribution $x_i \sim \mathcal{N}(0, \sigma^2/(1-\varphi^2))$. Assuming x_1 is drawn from



(a) MGDM, $T = 100$ (b) MGDM, $T = 36$ (c) MF–MGDM, $T = 157$

Figure 1: Densities of $\Phi(X)$, using fitted 2D Gaussians, at different stages of the descent for MGDM and MF–MGDM. In (b) and (c), T is the respective optimal number of steps to minimize KL divergence. The true distribution p is an AR(1) process with $\varphi = 0.1$ and $\sigma^2 = 0.99$.

this marginal, the likelihood is

$$p(x) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=2}^{d}(x_i - \varphi x_{i-1})^2 - \frac{1-\varphi^2}{2\sigma^2}x_1^2\right\}$$
$$\approx \exp\left\{\frac{\varphi}{\sigma^2}\sum_{i=2}^{d} x_i x_{i-1} - \frac{1+\varphi^2}{2\sigma^2}\sum_{i=1}^{d} x_i^2\right\}.$$

Thus, AR(1) is approximately an exponential family with the sufficient statistics

$$\Phi(x) = \left(\frac{1}{d}\sum_{i=2}^{d} x_i x_{i-1},\ \frac{1}{d}\sum_{i=1}^{d} x_i^2\right), \tag{7}$$

and is by the Boltzmann equivalence principle asymptotically equivalent with the microcanonical approximation with energy function (7).

Let us now approximate the microcanonical model using MGDM. We thus have an initial measure q_0 that is mapped through T steps of gradient descent to some final measure q_T. Figure 1a illustrates how the initial distribution in the energy space $\Phi_{\#}q_0$ is mapped to its final distribution $\Phi_{\#}q_T$ after $T = 100$ steps, bringing it close to the target energy. As can be seen in the pushforward of the true measure $\Phi_{\#}p$, however, true samples have a much greater variability in these statistics, making clear the need for regularization. If we instead stop the gradient descent earlier, after $T = 36$ steps, we obtain the distribution in Figure 1b, where we have preserved more of the entropy, but at the cost of a worse likelihood fit. The MF–MGDM, which we introduce below, performs well with respect to both aspects (Fig. 1c).

3.2 KL Divergence

Using the reverse KL divergence allows us to quantitatively analyze the method in examples like the AR(1) model where we have access to the density function of the target distribution. If ∇L is Lipschitz and the step size γ is smaller than the Lipschitz constant, the gradient step (4) is contractive and MGDM can be seen as a contractive residual flow. The log-likelihood $\log q_T$ is therefore

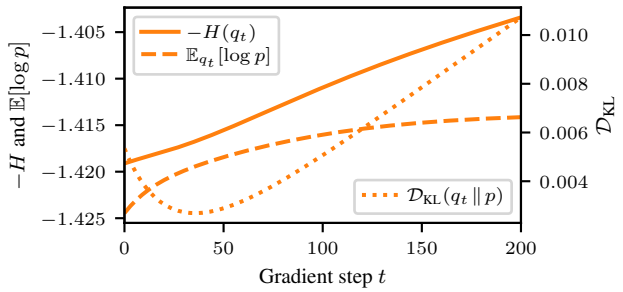
$$\log q_T(x) = \log q_0(z) - \sum_{t=0}^{T-1} \log|\det J_g(G_t(z))|, \tag{8}$$

where G_t denotes t compositions of g (with $G_0 \coloneqq \mathrm{I}$), and $z \coloneqq G_T^{-1}(x)$. The Jacobian $J_g(G_t(z))$ is computed by automatic differentiation through `torch.func` in PyTorch (Paszke et al., 2019, v2.1) (BSD-3). To arrive at the KL divergence, the expected values of (8) and $\log p$ in (1) are estimated by Monte Carlo.

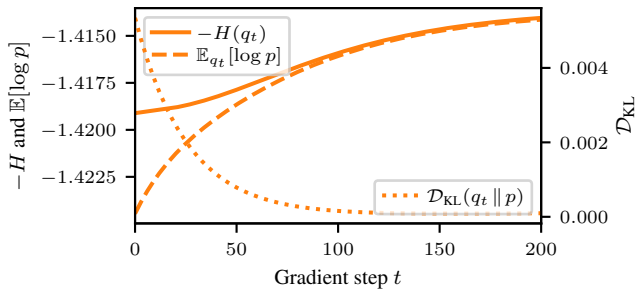
Going back to the AR(1) example, Figure 2a illustrates how the reverse KL divergence attains its minimum after $T = 36$ steps and then starts increasing; the improvement in likelihood fit gradually diminishes while the entropy keeps decreasing, causing an entropy collapse. In this case, the trade-off between entropy and log-likelihood is a false trade-off in that minimizing the KL divergence leaves us with a poor entropy *and* a poor expected log-likelihood, arguing against early stopping as a means of regularization. In contrast, we see that the proposed method, MF–MGDM, does not exhibit this problem in Figure 2b.

4 MEAN-FIELD MICROCANONICAL GRADIENT DESCENT

In the MGDM, the expected log-likelihood increases as the descent progresses and the energy approaches the target (assuming an appropriate energy function for the given distribution we model, e.g. the sufficient statistics as in the AR(1) case). Conversely, if too many iterates are performed, the energy vectors of the samples will be too close. Note that this happens even if the ε parameter of the microcanonical ensemble is chosen to be large, since the MGDM method will be concentrated over a small subset of Ω_ε. This observation leads to our



(a) MGDM



(b) MF–MGDM

Figure 2: Reverse KL divergence for the AR(1) example. The negative entropy and expected log-likelihood are plotted on the left-hand side, and the divergence on the right. Quantities are normalized by dimension.

proposition of the mean-field microcanonical gradient descent model (MF–MGDM).

4.1 The Model

(a) MGDM (b) MF–MGDM

Figure 3: Illustration of Φ-pushforward measures of the true distribution in blue centered close to the target energy α, and the approximation in orange. In the regular MGDM, each particle individually seeks to minimize its distance to the origin in energy space, potentially causing a collapse; in the mean-field version, the particles move approximately in parallel.

In the MF–MGDM, the mass of the initial distribution is pushed towards the target in energy space while attempting to reduce the collapse of the radius of the ball (or similarly the energy variance) and thereby reducing the entropy loss. The principle is illustrated in Figure 3.

Whereas the regular MGDM (Figure 3a) updates each sample *individually* with the objective of minimizing its energy distance (3) to the target, MF–MGDM (Figure 3b) updates several samples simultaneously so that they move towards the target energy *in aggregate*. Ultimately, the generated samples may not all be in Ω_ε depending on the choice of ε, but if necessary, this can be achieved by performing regular MGDM iterates after the MF–MGDM has converged.

Formally, define $\boldsymbol{x} = \{x^{(n)}\}_{n=1}^N \in \mathbb{R}^{Nd}$ as a collection of N particles, where a *particle* is a sample path in $\mathbb{R}^d$. Recalling the mean energy $\overline{\Phi}$ in (5), the new optimization objective is

$$\overline{L}(\boldsymbol{x}) \coloneqq \frac{N}{2}\|\overline{\Phi}(\boldsymbol{x}) - \alpha\|^2. \tag{9}$$

Denoting by $\mathcal{J}_\Phi(\boldsymbol{x})$ the concatenation of the Jacobians $J_\Phi(x^{(n)})$ of Φ with respect to each particle $x^{(n)}$,

$$\mathcal{J}_\Phi(\boldsymbol{x}) \coloneqq \begin{bmatrix} J_\Phi(x^{(1)}) & \cdots & J_\Phi(x^{(N)}) \end{bmatrix} \in \mathbb{R}^{K \times Nd}, \tag{10}$$

we define the mean-field gradient step as a gradient step for the objective (9), namely

$$\overline{g}(\boldsymbol{x}) \coloneqq \boldsymbol{x} - \gamma \mathcal{J}_\Phi^\top(\boldsymbol{x}) \left(\overline{\Phi}(\boldsymbol{x}) - \alpha\right). \tag{11}$$

The mean-field concept originates from statistical physics as a tool for studying macroscopic phenomena in large particle systems by averaging over microscopic interactions. In the context of game theory, for instance, mean-field games are multiagent problems where each agent has a negligible impact on the others, so that the dynamics of an agent depends on the law of the system. For an N-player system, the law is the empirical measure, for which a subclass of systems are those where the dynamics depend on the empirical mean. The mean-field limit is then when $N \to \infty$; see e.g. Carmona and Delarue (2018). We can think of (11) as corresponding to a discretization of a system of differential equations with mean-field interactions.

The MF–MGDM faces two challenges that the regular model does not. The first is that the sampling procedure requires simultaneous generation of multiple samples in order to compute $\overline{\Phi}$. This is solved efficiently by vectorizing the computation of $\mathcal{J}_\Phi(\boldsymbol{x})$ in (10). Furthermore, most applications call for generation of multiple samples, so the additional cost would be incurred at any rate by multiple invocations of MGDM.

The second challenge, which is inconsequential when only sampling, is that of computing the entropy, specifically computing the log-determinant of the Jacobian of a gradient step $\overline{g}$. The issue is that the samples are now coupled, resulting in the Jacobian being one large $Nd \times Nd$ matrix. Naively computing the determinant

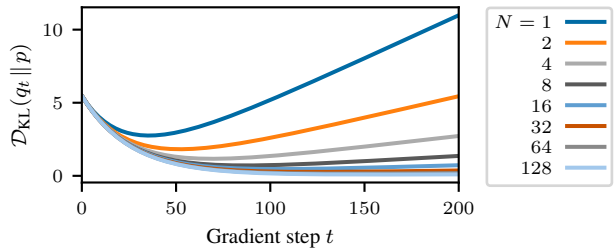


Figure 4: Reverse KL divergence through gradient descent with respect to the true model AR(1) for MF–MGDM with different mean-field batch sizes N, and with Monte Carlo sample size 128.

scales as $\mathcal{O}(N^3 d^3)$ (even keeping the Jacobian in memory is infeasible), but it is possible to rewrite it on a form that allows $\mathcal{O}(Nd^3)$ computation by writing the Jacobian as a sum of a block diagonal and a low-rank matrix, and then using the matrix determinant lemma (see Appendix A).

4.2 An Illustrative Example – Revisited

To demonstrate the effect of the mean-field gradient step, we return to the AR(1) example. Figure 1c shows the pushforward by Φ of the MF–MGDM approximation after 157 steps when minimum KL is achieved. We see now that the final distribution in energy space more closely aligns with that of the true measure, preventing the reduction of entropy observed in the MGDM (see Figure 1a). Tracking the reverse KL divergence for each gradient step, Figure 2b, we see an almost monotone decrease, avoiding the need for early stopping. If we break up the KL divergence into negative entropy and log-likelihood, we no longer observe an unbounded decrease in entropy. Instead, it stabilizes around a value close to the negative log-likelihood, resulting in a small KL divergence.

4.3 Theoretical Entropy Bound

We state a lower bound for the entropy rate, defined as the entropy normalized by the dimensionality of the space to which the signals belong. In MF–MGDM, the joint distribution is over N time series of length d, hence we have to normalize with Nd.

Theorem 4.1. *Assume $\Phi \in \mathbf{C}^2$, with β and η denoting the Lipschitz constants of Φ and $\nabla\Phi$, respectively. Denote $\overline{q}_T^N$ as the distribution of the MF–MGDM model with N particles after T iterations. Then the entropy rate $(Nd)^{-1}H(\overline{q}_T^N)$ admits, up to $\mathcal{O}(\gamma^2)$ terms, the lower bound*

$$(Nd)^{-1}H(\overline{q}_T^N) \geq (Nd)^{-1}H(\overline{q}_0^N) - 2\gamma\left(\eta\sqrt{K}\sum_{t=0}^{T-1}\mathbb{E}_{\overline{q}_t^N}\|\overline{\Phi}(\boldsymbol{X}) - \alpha\| + \frac{K}{Nd}\beta^2 T\right).$$

The entropy bound for the regular MGDM is recovered when $N = 1$ (since $\overline{\Phi}$ and Φ are then equal). Herein lies an explanation for the improvement in KL of the MF–MGDM. In both models, $\overline{\Phi}$ or Φ goes to α, whereby the cost in entropy for each gradient step is after a point mainly driven by the $\beta^2 T$-term, which can be made arbitrarily small in MF–MGDM by increasing N. This is also reflected empirically in Figure 4 where a monotonic improvement of KL divergence is observed as N grows. Note, however, that this is a lower bound, so it does not guarantee that MF–MGDM always preserves entropy better than MGDM (although this is observed numerically), but it does provide a better guarantee. The proof of Theorem 4.1 is given in Appendix B.

5 NUMERICAL EXPERIMENTS

To evaluate the performance of this sampling scheme, we apply it to time series and a two-dimensional texture in the form of bubbles. In Appendix E, we also provide results for two-dimensional Ising processes. In all experiments except for CIR data, the sampling processes are initialized with Gaussian white noise. For CIR, it is instead truncated Gaussian or exponential white noise (see details in Appendix C).

5.1 Time Series

To compare the different approximation models on synthetic data, we use time series models that have density functions in closed form, allowing for evaluation of the reverse KL divergence. We generate 10 000 samples of length 1 024 and take the average energy over these samples as target energy, to simulate the idealized setting where the true energy vector is known, avoiding bias. The KL divergence is estimated by generating 128 samples from the respective models and recording the divergence after each gradient step.

We used the following energy functions:

a. AR(1) approximate sufficient statistics (7) (or equivalently, autocovariance at lags 0 and 1);

b. First moments of the second-order scattering transform (with complex modulus as nonlinearity), using filters from the Kymatio package (Andreux et al., 2020, v0.3) (BSD-3);

Table 1: Minimum reverse KL divergence over T for different distributions and approximation models, where Reg. denotes the regular MGDM whereas MF is the mean-field version; $N = 128$.

	ACF eqn. (7)		ScatMean		ScatCov		ScatSpectra	
	Reg.	MF	Reg.	MF	Reg.	MF	Reg.	MF
AR(0.1)	2.76	**0.09**	4.24	1.99	5.47	4.04	5.44	2.32
AR$(0.2, -0.1)$	9.44	**3.81**	17.98	10.55	25.91	14.84	27.33	9.60
AR$(-0.1, 0.2, 0.1)$	30.04	26.39	20.98	15.18	29.55	18.01	28.46	**10.13**
CIR$(1/2, 1, 1)$	219.40	214.65	170.99	168.88	121.17	59.21	105.05	**18.56**
CIR$(1/\sqrt{2}, \sqrt{2}, 1)$	104.49	**87.96**	182.32	179.34	223.63	204.79	203.46	201.44

c. Second moments of the second-order scattering transform, augmented with filters shifted by 0 and $\pi/3$ in the first-order coefficients, and using ReLU of the real part as nonlinearity. Finally, we perform a dimensionality reduction by using principal component analysis (PCA) on transforms applied to Gaussian white noise;

d. Scattering spectra from Morel et al. (2023) (MIT License), taking the modulus of those coefficients which are complex, and thereby ignoring the phase.

These energy functions are applied to two types of synthetic data: autoregressive models of order p (AR(p)) models and Cox–Ingersoll–Ross (CIR) models.

AR(p) An AR(p) model with parameters $\varphi_1, \dots, \varphi_p$ and σ is a generalization of the AR(1) process in (6) and is defined by the recursion $x_i = \sum_{j=1}^{p} \varphi_j x_{i-j} + \sigma\varepsilon_i$, with white noise $(\varepsilon_i)_i$, and is stationary if the roots of the characteristic polynomial $\pi(z) = 1 - \sum_j \varphi_j z^j$ are outside the unit circle. In Table 1, the models are denoted AR$(\varphi_1, \dots, \varphi_p)$, and σ is chosen as to obtain unit marginal variance.

CIR The CIR model (Cox et al., 1985) is a diffusion process that is commonly used for modeling short-term interest rates. It is related to the Ornstein–Uhlenbeck process, which can be seen as a continuous version of AR(1), but differs in the way that the diffusion term is scaled by the square root of the rate $r \in \mathbb{R}^+$ to give

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t,$$

where W is a Brownian motion. The process admits a stationary distribution, and the distribution at time t given the value at an earlier time $s < t$ is a scaled non-central χ^2 distribution which can be written in closed form, allowing for explicit evaluation of the likelihood of a discretization in an autoregressive fashion. The distribution of r_0 can be taken to be the marginal distribution, i.e., a gamma distribution. In the experiments, we use the discretization $\Delta t = 1$, and the models are identified as CIR(κ, θ, σ). The CIR process is non-negative, so projected gradient descent, described in Appendix C, has to be used when approximating this distribution in the context of MGDM.

Results For each model and energy function, the reverse KL divergence was computed at each step through the descent. The minimum divergence achieved is displayed in Table 1. For every true distribution, we present results also for energy functions that are not necessarily a good choice, given the true model. We see here that MF–MGDM consistently outperforms MGDM.

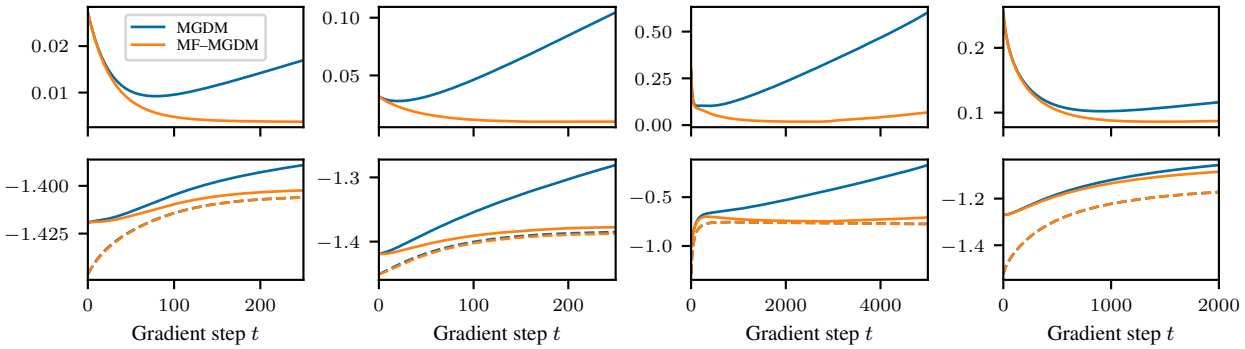
The KL divergence through the descent as a function of iteration number is shown in Figure 5, as well as its constituents entropy and expected log-likelihood. Here we have only plotted results for the energy function that best approximates each distribution in accordance with Table 1. Here we see again that the mean-field model retains more entropy, and the difference is marginal between the expected likelihoods of the two models.

Another important difference here is that while MGDM needs to be stopped early to prevent the entropy from collapsing, this is not the case for MF–MGDM. Indeed, we see that the entropy stabilizes after a certain number of steps similarly to the log-likelihood. This is important because in a real-world setting, the true distribution is not known and the reverse KL divergence is not computable, so we cannot reasonably estimate the number of gradient steps to perform in order to balance the entropy loss with the increase of expected log-likelihood. For MF–MGDM, we can run the sampling until convergence while being less sensitive to this type of overfitting.

5.2 Bubbles

We also compare the methods for images in the form of a two-dimensional bubble texture[1] used by Zhang and Mallat (2021) for illustrating the expressiveness of their energy function consisting of *phase harmonic covariances*. It is treated in more detail in Appendix D, but can be described as a variant of the scattering

[1]Downloaded from `https://cloud.irit.fr/s/PIVBO4JJBT73rcp`



(a) AR$(0.2, -0.1)$ (b) AR$(-0.1, 0.2, 0.1)$ (c) CIR$(1/2, 1, 1)$ (d) CIR$(1/\sqrt{2}, \sqrt{2}, 1)$

Figure 5: Reverse KL divergence (top), negative entropy (bottom, solid) and log-likelihood (bottom, dashed) through the descent. Blue is regular MGDM and orange is MF–MGDM. The energy function used for each distribution is the corresponding optimal energy function according to Table 1, i.e., (a) and (d) use ACF while (b) and (c) use scattering spectra. $N = 128$. Quantities are normalized by dimension.

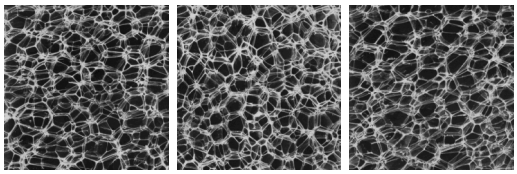
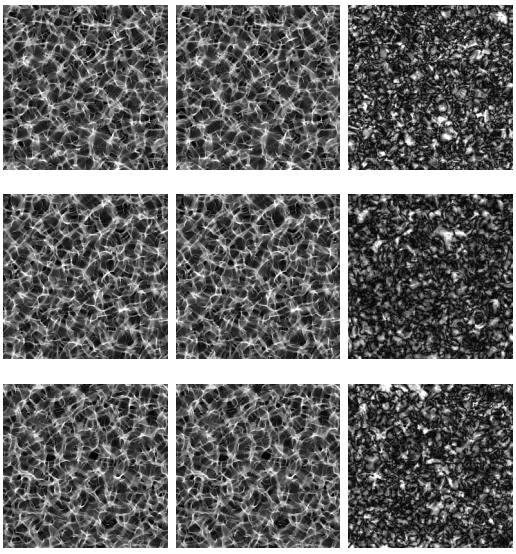


Figure 6: True samples of 256×256 bubbles texture.

transform, computing second moments of the wavelet transform with different phase shifts. We recreate their results here, and compare them with the MF–MGDM using the same energy function.

To the eye, as seen in Figure 7, the difference between samples from the two models, given the same initialization, is quite small. We therefore include the absolute value of the pixel-wise difference in terms of the marginal pixel-value standard deviation of the true sample.

To quantitatively compare the models, we compute their resulting entropies through the descent on 32×32 images. The target energy is defined as the average energy over all true samples, which are disjoint patches of four 256 true samples. The entropies are displayed in Figure 9, showing that the initial entropy is better preserved for MF–MGDM, and samples are shown in Figure 8. To make the models converge in fewer steps while avoiding big updates during the first iterations, a variable step length of piecewise exponentially increasing values is used. This is the reason why the difference appears only after 300 iterations.



(a) Reg (b) MF (c) Diff

Figure 7: Generated samples of 256×256 bubbles texture conditioned on a single true sample. The first column Reg consists of samples from the regular MGDM; the second is the mean-field version. Both samples come from the same Gaussian white noise initialization in each row, explaining the similarity between these columns. The third column displays the absolute value of the difference relative to one standard deviation of the marginal distribution of the true sample, with zero being black and white being the 99^{th} percentile of all pixel differences across all generated samples (in this case 0.22).

6 LIMITATIONS

First, we emphasize the ergodicity assumption of the signals. Next, the MGDM requires the energy function Φ to be differentiable so it is not straightforward to include e.g. order statistics constraints. As far as we know, there is presently no modification of the MGDM which allows for a stable way of inverting the descent

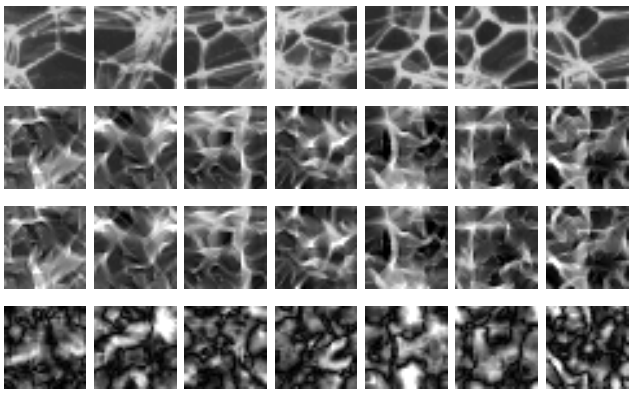


Figure 8: Realizations and approximations of 32×32 bubble textures. From top to bottom: True patches, MGDM, MF–MGDM, relative difference between MGDM and MF–MGDM. The difference is defined as in Fig 7, with black at 0 and white at 0.24. Except the top row, seeds are shared column-wise.

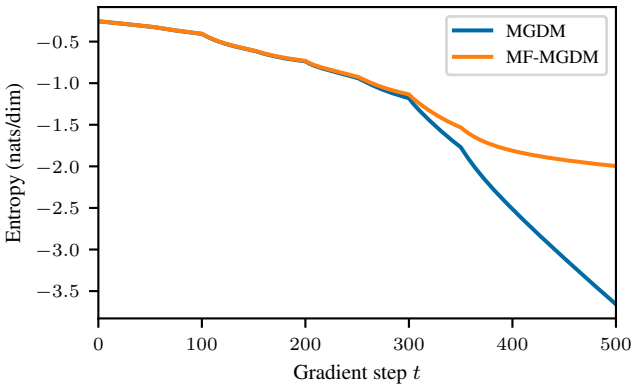


Figure 9: Differential entropies for MGDM approximations of a 32×32 bubble texture. A variable step size γ has been used, which is why the graphs diverge relatively late compared e.g. to Figure 5.

in order to be able to compute forward KL in the usual case where the true distribution is not known, forcing only qualitative evaluation of performance on real-world data where the true likelihood function is not known. Finally, although the width ε of Ω_ε is important for a good KL fit, exactly how to tune this parameter is left for future work.

7 CONCLUSIONS

The MGDM provides efficient sampling of high-dimensional distributions, but can suffer from a significant loss of entropy. Propagating too far into the descent is shown to overfit to the target energy that the model is conditioned on, meaning that the variance of the energy for the model is much too small as for what to expect from true distributions. Regularizing by early stopping in the descent mitigates this issue somewhat, but at the price of a worse fit to the true distribution and a larger bias from the initial distribution. The mean-field regularization of the model in the form of MF–MGDM leverages parallel sampling to mitigate the problem, improving the rate at which entropy is lost without a significant impact on the likelihood fit. Future work will explore better initial distributions and more sophisticated update steps. These will in turn open the door to considering forward KL divergence metrics, removing the need for access to the likelihood of the target distribution.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

E. Allys, T. Marchand, J.-F. Cardoso, F. Villaescusa-Navarro, S. Ho, and S. Mallat. New interpretable statistics for large-scale structure analysis and generation. *Physical Review D*, 102(10):103506, 2020.

M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. Chaudhary, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, and M. Eickenberg. Kymatio: Scattering transforms in Python. *Journal of Machine Learning Research*, 21 (60):1–6, 2020.

J. Andén, V. Lostanlen, and S. Mallat. Joint time–frequency scattering. *IEEE Transactions on Signal Processing*, 67(14):3704–3718, 2019.

C. Auclair, E. Allys, F. Boulanger, M. Béthermin, A. Gkogkou, G. Lagache, A. Marchal, M.-A. Miville-Deschênes, B. Régaldo-Saint Blancard, and P. Richard. Separation of dust emission from the cosmic infrared background in Herschel observations with wavelet phase harmonics. *Astronomy & Astrophysics*, 681:A1, 2023.

A. Brochard, B. Błaszczyszyn, S. Zhang, and S. Mallat. Particle gradient descent model for point process generation. *Statistics and Computing*, 32(3):49, 2022a.

A. Brochard, S. Zhang, and S. Mallat. Generalized rectifier wavelet covariance models for texture synthesis. In *International Conference on Learning Representations*, 2022b.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

J. Bruna and S. Mallat. Multiscale sparse microcanonical models. *Mathematical Statistics and Learning*, 1 (3):257–315, 2019.

H. Buehler, B. Horvath, T. Lyons, I. Perez Arribas, and B. Wood. A data-driven market simulator for small data environments. *arXiv preprint arXiv:2006.14498*, 2020.

R. Carmona and F. Delarue. *Probabilistic theory of mean field games with applications I-II*. Springer, 2018.

S. Cheng, R. Morel, E. Allys, B. Ménard, and S. Mallat. Scattering spectra models for physics. *PNAS nexus*, 3(4), 2024.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

V. Chudáček, J. Andén, S. Mallat, P. Abry, and M. Doret. Scattering transform for intrapartum fetal heart rate characterization and acidosis detection. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2898–2901. IEEE, 2013a.

V. Chudáček, J. Andén, S. Mallat, P. Abry, and M. Doret. Scattering transform for intrapartum fetal heart rate variability fractal analysis: a case-control study. *IEEE Transactions on Biomedical Engineering*, 61(4):1100–1108, 2013b.

V. Chudáček, R. Talmon, J. Andén, S. Mallat, R. R. Coifman, P. Abry, and M. Doret. Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6373–6376, 2014.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2nd edition, 2006. ISBN 978-0-471-24195-9.

J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385, 1985. doi: 10.2307/1911242.

D. Dowson and A. Wragg. Maximum-entropy distributions having prescribed first and second moments. *IEEE Transactions on Information Theory*, 19(5): 689–693, 1973.

Y. Du and I. Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat. Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3D electronic densities. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

M. Eickenberg, E. Allys, A. M. Dizgah, P. Lemos, E. Massara, M. Abidi, C. Hahn, S. Hassan, B. R.-S. Blancard, S. Ho, et al. Wavelet moments for cosmological parameter estimation. *arXiv preprint arXiv:2204.07646*, 2022.

R. S. Ellis, K. Haven, and B. Turkington. Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles. *Journal of Statistical Physics*, 101(5):999–1064, 2000.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

O. E. Lanford. Time evolution of large classical systems. In *Dynamical Systems, Theory and Applications*, pages 1–111. Springer, Berlin, Heidelberg, 1975.

W. Lenz. Beiträge zum Verständnis der magnetischen Eigenschaften in festen Körpern. *Physikalische Zeitschrift*, 21:613–615, 1920.

R. Leonarduzzi, G. Rochette, J.-P. Bouchaud, and S. Mallat. Maximum-entropy scattering models for financial time series. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5496–5500, 2019.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

V. Lostanlen, G. Lafay, J. Andén, and M. Lagrange. Relevance-based quantization of scattering features for unsupervised mining of environmental audio. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1):1–10, 2018.

V. Lostanlen, C. El-Hajj, M. Rossignol, G. Lafay, J. Andén, and M. Lagrange. Time–frequency scattering accurately models auditory similarities between instrumental playing techniques. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):3, 2021.

T. Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 2016.

S. Mallat, S. Zhang, and G. Rochette. Phase harmonic correlations and convolutional neural networks. *Information and Inference: A Journal of the IMA*, 9 (3):721–747, 2019.

R. Morel, G. Rochette, R. Leonarduzzi, J.-P. Bouchaud, and S. Mallat. Scale dependencies and self-similar models with wavelet scattering spectra. *Available at SSRN 4516767*, 2023.

H. Ni, L. Szpruch, M. Sabate-Vidales, B. Xiao, M. Wiese, and S. Liao. Sig-Wasserstein GANs for time series generation. *arXiv preprint arXiv:2111.01207*, 2021.

A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

E. Oyallon, S. Zagoruyko, G. Huang, N. Komodakis, S. Lacoste-Julien, M. Blaschko, and E. Belilovsky. Scattering networks for hybrid representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2208–2221, 2019.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

A. Strominger. Microcanonical quantum field theory. *Annals of Physics*, 146(2):419–457, 1983.

H. Touchette. Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels. *Journal of Statistical Physics*, 159(5):987–1016, 2015.

P. Villoutreix, J. Andén, B. Lim, H. Lu, I. G. Kevrekidis, A. Singer, and S. Y. Shvartsman. Synthesizing developmental trajectories. *PLoS computational biology*, 13(9):e1005742, 2017.

P. A. Warrick, V. Lostanlen, M. Eickenberg, J. Andén, and M. N. Homsi. Arrhythmia classification of 12-lead electrocardiograms by hybrid scattering-LSTM networks. In *2020 Computing in Cardiology*, pages 1–4, 2020.

L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

S. Zhang and S. Mallat. Maximum entropy models from phase harmonic covariances. *Applied and Computational Harmonic Analysis*, 53:199–230, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, Section 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, e.g. top left p. 6]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, provided in supplemental material]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes, in appendix]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [No, KL computations too heavy to provide error bars]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, stated in supplemental material]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Appendix – Computing the Jacobian Determinant in MF–MGDM

Without loss of generality, assume $\alpha = 0$ (otherwise, we simply redefine $\Phi(x)$ to be $\Phi(x) - \alpha$). Denote $\overline{g}^{(n)}$ as the update corresponding to particle $x^{(n)}$:

$$\overline{g}^{(n)}(\boldsymbol{x}) = x^{(n)} - \gamma \sum_{k=1}^{K} \nabla \Phi_k(x^{(n)}) \overline{\Phi}_k(\boldsymbol{x}).$$

Then the Jacobian w.r.t. a possibly different particle $x^{(m)}$ is, stated by index,

$$\begin{aligned}
\partial_{x_j^{(m)}} \overline{g}_i^{(n)}(\boldsymbol{x}) &= \delta_{m,n}\delta_{i,j} - \gamma \sum_k \partial_{x_j^{(m)}} \left(\partial_{x_i^{(n)}} \Phi_k(x^{(n)}) \cdot \overline{\Phi}_k(\boldsymbol{x}) \right) \\
&= \delta_{m,n}\delta_{i,j} - \gamma \sum_k \left(\delta_{m,n} \partial_{x_j^{(n)}} \partial_{x_i^{(n)}} \Phi_k(x^{(n)}) \cdot \overline{\Phi}_k(\boldsymbol{x}) + \frac{1}{N} \partial_{x_i^{(n)}} \Phi_k(x^{(n)}) \cdot \partial_{x_j^{(m)}} \Phi_k(x^{(m)}) \right),
\end{aligned}$$

or, stated by block,

$$J_{\overline{g}^{(n)}}(x^{(m)}) = \delta_{m,n} \cdot \left(\mathrm{I}_d - \gamma \sum_k H_{\Phi_k}(x^{(n)}) \overline{\Phi}_k(\boldsymbol{x}) \right) - \frac{\gamma}{N} J_\Phi^\top(x^{(n)}) J_\Phi(x^{(m)}),$$

where I_d is the $d \times d$ identity matrix. Recall the concatenation (10) of the Jacobians,

$$\mathcal{J}_\Phi(\boldsymbol{x}) = \begin{bmatrix} J_\Phi(x^{(1)}) & \cdots & J_\Phi(x^{(N)}) \end{bmatrix},$$

and define the block-diagonal matrix

$$\mathcal{H}_{\Phi_k}(\boldsymbol{x}) = \operatorname{diag}\left\{ H_{\Phi_k}(x^{(n)}) \right\}_{n=1}^{N} = \begin{bmatrix} H_{\Phi_k}(x^{(1)}) & & \\ & \ddots & \\ & & H_{\Phi_k}(x^{(N)}) \end{bmatrix}. \tag{12}$$

Then, the entire Jacobian of $\overline{g}$ can be expressed as

$$J_{\overline{g}}(\boldsymbol{x}) = \mathrm{I}_{Nd} - \gamma \sum_k \mathcal{H}_{\Phi_k}(\boldsymbol{x}) \overline{\Phi}_k(\boldsymbol{x}) - \frac{\gamma}{N} \mathcal{J}_\Phi^\top(\boldsymbol{x}) \mathcal{J}_\Phi(\boldsymbol{x}). \tag{13}$$

Using the matrix determinant lemma, and that

$$\mathrm{I}_{Nd} - \gamma \sum_k \mathcal{H}_{\Phi_k} \overline{\Phi}_k$$

is block-diagonal (and thereby also its inverse), the determinant can be reformulated as

$$\begin{aligned}
\det J_{\overline{g}} &= \det \left(\mathrm{I}_{Nd} - \gamma \sum_k \mathcal{H}_{\Phi_k} \overline{\Phi}_k - \frac{\gamma}{N} \mathcal{J}_\Phi^\top \mathcal{J}_\Phi \right) \\
&= \det \left(\mathrm{I}_{Nd} - \gamma \sum_k \mathcal{H}_{\Phi_k} \overline{\Phi}_k \right) \det \left(\mathrm{I}_K - \frac{\gamma}{N} \mathcal{J}_\Phi \left(\mathrm{I}_{Nd} - \gamma \sum_k \mathcal{H}_{\Phi_k} \overline{\Phi}_k \right)^{-1} \mathcal{J}_\Phi^\top \right) \\
&= \det \operatorname{diag} \left\{ \left(\mathrm{I}_d - \gamma \sum_k H_{\Phi_k}^{(n)} \overline{\Phi}_k \right) \right\}_n \det \left(\mathrm{I}_K - \frac{\gamma}{N} \mathcal{J}_\Phi \operatorname{diag} \left\{ \left(\mathrm{I}_d - \gamma \sum_k H_{\Phi_k}^{(n)} \overline{\Phi}_k \right)^{-1} \right\}_n \mathcal{J}_\Phi^\top \right) \\
&= \prod_n \det \left(\mathrm{I}_d - \gamma \sum_k H_{\Phi_k}^{(n)} \overline{\Phi}_k \right) \det \left(\mathrm{I}_K - \gamma \frac{1}{N} \sum_n J_\Phi^{(n)} \left(\mathrm{I}_d - \gamma \sum_k H_{\Phi_k}^{(n)} \overline{\Phi}_k \right)^{-1} \left(J_\Phi^{(n)} \right)^\top \right).
\end{aligned}$$

B Appendix – Proof of Theorem 4.1

As in previous Section A, we assume without loss of generality that $\alpha = 0$.

From (8) we get

$$H(\overline{q}_T^N) = -\mathbb{E}_{\overline{q}_T^N}[\log \overline{q}_T^N(\boldsymbol{X})] = -\mathbb{E}_{\overline{q}_0^N}\left[\log \overline{q}_0^N(\boldsymbol{X}) - \sum_{t=0}^{T-1} \log|\det J_{\overline{g}}(\overline{g}_t(\boldsymbol{X}))|\right]$$
$$= H(\overline{q}_0^N) + \sum_{t=0}^{T-1} \mathbb{E}_{\overline{q}_t^N}[\log|\det J_{\overline{g}}(\boldsymbol{X})|.] \tag{14}$$

so we want to lower-bound $\log|\det J_{\overline{g}}|$. By (13) we see that we can write $J_{\overline{g}}(\boldsymbol{x})$ on the form $\mathrm{I} - \gamma A$. We have

$$\frac{d}{d\gamma}\det(\mathrm{I} - \gamma A)\bigg|_{\gamma=0} = -\det(\mathrm{I})\,\mathrm{Tr}(\mathrm{I}^{-1}A) = -\,\mathrm{Tr}\,A,$$

which yields the Taylor approximation

$$\det(\mathrm{I} - \gamma A) = 1 - \gamma\,\mathrm{Tr}\,A + \mathcal{O}(\gamma^2).$$

This, together with the lower bound for the logarithm

$$\log(1-x) \geq -2x$$

for $x \in [0, \frac{3}{4}]$, results in the lower bound (suppressing the argument $(\boldsymbol{x})$)

$$\log|\det J_{\overline{g}}| \geq -2\gamma\left|\mathrm{Tr}\left(\sum_k \mathcal{H}_{\Phi_k}\overline{\Phi}_k + \frac{1}{N}\mathcal{J}_\Phi^\top\mathcal{J}_\Phi\right)\right| + \mathcal{O}(\gamma^2) \tag{15}$$

for γ small enough. Thus, we seek an upper bound to

$$\left|\mathrm{Tr}\left(\sum_k \mathcal{H}_{\Phi_k}\overline{\Phi}_k + \frac{1}{N}\mathcal{J}_\Phi^\top\mathcal{J}_\Phi\right)\right| \leq \sum_k |\,\mathrm{Tr}(\mathcal{H}_{\Phi_k})\overline{\Phi}_k| + \frac{1}{N}|\,\mathrm{Tr}(\mathcal{J}_\Phi^\top\mathcal{J}_\Phi)|. \tag{16}$$

The Lipschitz assumption on Φ yields $\|J_\Phi(x)\|_2 \leq \beta$ for all x, so that for any particle (here suppressing the argument $(x^{(i)})$),

$$\mathrm{Tr}(J_\Phi^\top J_\Phi) = \mathrm{Tr}(J_\Phi J_\Phi^\top) = \sum_k \lambda_k(J_\Phi J_\Phi^\top) \leq K\lambda_{\max}(J_\Phi J_\Phi^\top) = K\|J_\Phi^\top\|_2^2 = K\|J_\Phi\|_2^2 \leq K\beta^2,$$

whereby the second term of (16) becomes

$$\frac{1}{N}\,\mathrm{Tr}(\mathcal{J}_\Phi^\top(\boldsymbol{x})\mathcal{J}_\Phi(\boldsymbol{x})) = \frac{1}{N}\sum_{i=1}^N \mathrm{Tr}(J_\Phi^\top(x^{(i)})J_\Phi(x^{(i)})) \leq \frac{1}{N}NK\beta^2 = K\beta^2. \tag{17}$$

Similarly, the Lipschitz assumption on $\nabla\Phi$ together with symmetry of H implies $\|H_{\Phi_k}(x)\|_2 = |\lambda|_{\max}(H_{\Phi_k}(x)) \leq \eta$ for all k and x, and in turn,

$$\sum_k |\,\mathrm{Tr}(H_{\Phi_k})\overline{\Phi}_k| \leq \sum_k d|\lambda|_{\max}(H_{\Phi_k})|\overline{\Phi}_k| \leq d\eta\|\overline{\Phi}\|_1 \leq d\eta\sqrt{K}\|\overline{\Phi}\|_2.$$

Thus, the first term of (16) becomes

$$\sum_k |\,\mathrm{Tr}(\mathcal{H}_{\Phi_k}(\boldsymbol{x}))\overline{\Phi}_k(\boldsymbol{x})| = \sum_k \left|\sum_{i=1}^N \mathrm{Tr}(H_{\Phi_k}(x^{(i)}))\overline{\Phi}_k(\boldsymbol{x})\right| \leq Nd\eta\sqrt{K}\|\overline{\Phi}(\boldsymbol{x})\|_2. \tag{18}$$

Inserting (17) and (18) into (16), we see that the $\log|\det J_{\overline{g}}|$ bound (15) becomes

$$\log|\det J_{\overline{g}}(\boldsymbol{x})| \geq -2\gamma\left(Nd\eta\sqrt{K}\|\overline{\Phi}(\boldsymbol{x})\|_2 + K\beta^2\right) + \mathcal{O}(\gamma^2).$$

Hence, the lower bound on the entropy rate, up to second order terms in γ, becomes

$$(Nd)^{-1}H(\overline{q}_T^N) \geq (Nd)^{-1}H(\overline{q}_0^N) - 2\gamma\left(\eta\sqrt{K}\sum_{t=0}^{T-1}\mathbb{E}_{\overline{q}_t^N}\|\overline{\Phi}(\boldsymbol{X})\|_2 + \frac{K}{Nd}\beta^2 T\right).$$

C Appendix – Projected Gradient Descent

In the projected gradient descent used for CIR models, the generating procedure is to update the sample according to the gradient steps while satisfying the constraint of remaining in the positive cone $x \geq 0$. A basic implementation is to alternate between a gradient step and a projection step, where the updated sample is projected onto the feasible set, which in practice amounts to applying a ReLU to the sample after each step; let $g : \mathcal{X} \to \mathcal{X}$ denote the gradient update (regular or mean-field) and $\underline{g}$ the projected gradient update, then

$$\underline{g} = \mathrm{ReLU} \circ g.$$

The problem with this definition is that the Jacobian becomes singular if an update is masked by the ReLU, resulting in the determinant being zero. Therefore, we instead use the update

$$\underline{g}_i(x) = \begin{cases} g_i(x), & g_i(x) \geq 0, \\ x_i, & g_i(x) < 0. \end{cases}$$

Hence, if a component in the sample is negative after the gradient step g, it is replaced by its prior value. In this case, the Jacobian determinant is the same as only looking at the components of the sample that have been updated.

Another aspect of the projected version of MGDM is the choice of initial measure. If the support of the marginal distribution is all of $\mathbb{R}$, the maximum entropy distribution conditioned on the first two moments is the Gaussian. Thus, in this case, the MGDM is initialized with Gaussian white noise. For the CIR process, the support of the marginal distribution is $\mathbb{R}^+$, and, given that it exists, the corresponding maximum entropy distribution is either the exponential (if the mean and standard deviation are equal) or the truncated Gaussian (Dowson and Wragg, 1973).

D Appendix – Energy Functions

In this section, the energy functions used in Section 5 of the main text are presented in more detail, beginning with a brief introduction to the scattering transform proposed by Mallat (2012).

D.1 Scattering Transform

The scattering transform is a feature transform designed to incorporate invariances and stabilities that the data is known or assumed to satisfy, using tools from signal processing. The primary tool is the wavelet transform, which is a time-frequency decomposition of the signal.

A wavelet ψ is a pulse-like oscillation integrating to zero. It comes in many different forms, an example being the Morlet wavelet which is an offset complex exponential $e^{i\omega t} - \kappa_\omega$ times a Gaussian envelope, or the Haar wavelet

$$\psi(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2}, \\ -1, & \frac{1}{2} \leq t < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The wavelet acts as a bandpass filter, concentrating at a certain frequency in the Fourier domain. Starting with a so-called mother wavelet ψ, a filter bank $\{\psi_\lambda\}_\lambda$ is created by scaling the mother wavelet as $\psi_\lambda(t) = \lambda\psi(\lambda t)$, typically

accompanied by a low-pass filter ϕ, covering the frequency spectrum. In this work, $\lambda = 2^0, 2^{-1}, \ldots, 2^{-(J-1)}$, i.e., one wavelet per octave. Abusing the notation, we denote the lowpass filter by $\psi_{\lambda_J} := \phi$, and define the wavelet transform W as convolutions with the lowpass and wavelet filters:

$$Wx := \{x * \psi_{\lambda_j}\}_j.$$

In practice, the low-frequency filters for which the scale 2^j exceeds the signal length are replaced with fixed bandwidth filters, as implemented in the `kymatio` library (Andreux et al., 2020, v0.3).

Denoting $\hat{\psi}$ as the Fourier transform of ψ, observe that $\hat{\psi_\lambda}(\omega) = \hat{\psi}(\frac{\omega}{\lambda})$. Thus, not only does the central frequency shift, but the bandwidth is by construction also scaled, with wider bandwidth for higher frequencies and vice versa. This enables the Lipschitz continuity of the final transform with respect to deformations (Mallat, 2012). To incorporate translation invariance, the features are averaged over time, either by integrating over the entire signal or convolving with a low-pass filter to obtain local invariance. Since $\int \psi(t)dt = 0$, however, a nonlinearity must be applied before averaging. This is typically the complex modulus, $|W|$. The first-order scattering coefficients are thus defined as

$$S_1 x := \int |W|x(t)dt = \left\{ \int |x * \psi_{\lambda_j}|(t)dt \right\}_j,$$

where the average is replaced by a convolution with the low-pass filter ϕ to the windowed version. Note that S_1 can be seen as a layer of a convolutional neural network with predetermined weights and average pooling.

When averaging over time, high-frequency information is lost. To recover some of this information, a frequency decomposition by wavelet transform can be performed again on the first-order wavelet modulus $|x * \psi_\lambda|$ yielding second-order coefficients

$$S_2 x := \int |W| \circ |W|x(t)dt = \left\{ \int \big| |x * \psi_{\lambda_j}| * \psi_{\lambda_{j'}} \big|(t)dt \right\}_{j,j'}.$$

Although not used here, higher orders can be defined similarly, corresponding to a deeper CNN. The scattering transform of order two is finally defined as the collection of the coefficients S_1 and S_2,

$$Sx := S_1 x \cup S_2 x. \tag{19}$$

D.2 Energy Functions Used in Time Series Experiments

The energy functions used in the synthetic time series experiments of Section 5.1 are now defined, using the same enumeration as above.

a. AR(1) sufficient statistics are defined in (7) but are here restated for completeness:

$$\Phi(x) = \left(\frac{1}{d} \sum_{i=2}^{d} x_i x_{i-1},\ \frac{1}{d} \sum_{i=1}^{d} x_i^2 \right).$$

b. Scattering transform as defined in (19). The filters used are Morlet wavelets created using `kymatio`'s (Andreux et al., 2020, v0.3) filter factory. The maximum scale 2^J is set to J=8, but only the four and six highest frequency wavelets are used for the first and second order coefficients, respectively, in order to reduce dimensionality of Φ and remove coefficients that carry little information for the AR processes.

c. A primal version of phase harmonic correlations in (Mallat et al., 2019). This replaces the phase modulus with a ReLU ρ of the real part as nonlinearity, preserving some information on the phase φ:

$$U_1[\lambda]x := \rho(\text{Real}(x * \psi_\lambda)) = |x * \psi_\lambda|\, \rho\big(\cos\varphi(x * \psi_\lambda)\big) = |x * \psi_\lambda|\, \mathbb{1}\left\{ \varphi(x * \psi_\lambda) \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right] \right\}.$$

By shifting the phase with α, we retain even more phase information,

$$U_1[\lambda, \alpha]x := \rho(\text{Real}(e^{-i\alpha} x * \psi_\lambda)) = |x * \psi_\lambda|\, \mathbb{1}\left\{ \varphi(x * \psi_\lambda) \in \left[-\frac{\pi}{2} + \alpha, \frac{\pi}{2} + \alpha \right] \right\}.$$

The second order operators are defined in a similar cascading fashion as the scattering transform. We use the same filters as in b., and let $\alpha \in \{0, \pi/3\}$ in the first order and $\alpha = 0$ in the second. Instead of simply averaging to achieve translation invariance, the second central moments are computed. Many of these covariances are negligible, resulting in a large, sparse feature vector. To reduce dimensionality, we use PCA based on Gaussian white noise as an approximation.

d. Scattering spectra by Morel et al. (2023) consists of

$$\Phi_1 x := S_1 x = \int |W|x(t)dt,$$

together with a sparse approximation of the cross-correlations between and autocorrelations among W and $|W|$. Assumptions on the power spectrum together with small filter overlap in frequency yield an approximation of first block with correlations between Wx and $Wx(\cdot - \tau)$ consisting of the diagonal elements

$$\Phi_2 x := \int |W|^2 x(t)dt = \left\{ \int |x * \psi_{\lambda_j}|^2(t)dt \right\}_j,$$

denoted the *wavelet spectrum*. The second block of covariances between W and $|W|$ is also sparse, being negligible for $\tau = 0$. Furthermore, due to the modulus shifting the frequency band to zero, correlations with Wx for higher frequency filters can be disregarded, resulting in

$$\Phi_3 x := \left\{ \int x * \psi_{\lambda_{j'}}(t) \left| x * \psi_{\lambda_j} \right|(t)dt \right\}_{j \leq j'}.$$

The third and final block is that of the correlations between $|W|x$ and $|W|x(\cdot - \tau)$. This is not sparse, so another wavelet transform is applied, giving the second order transform $W \circ |W|$. Similar arguments as above can be made for sparsity, and the following coefficients are selected:

$$\Phi_4 x = \left\{ \int \left| x * \psi_{\lambda_j} \right| * \psi_{\lambda_{j''}}(t) \left| x * \psi_{\lambda_{j'}} \right| * \psi^*_{\lambda_{j''}}(t)dt \right\}_{j \leq j',\, \max(j,j') < j''},$$

where ψ^* is the complex conjugate of ψ.

Finally, the coefficients are normalized according to the wavelet spectrum of the observed data.

In the experiments, the package of Morel et al. (2023) available at `https://github.com/RudyMorel/scattering_spectra` is used, with $J = 7$ scales and subsequent parameters set to default values. In order to further reduce feature dimensionality, the modulus was applied on the complex components.

D.3 Energy Function Used in 2D Bubbles Texture Experiments

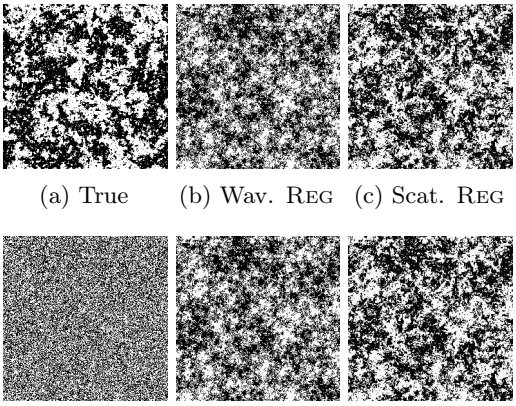
In the case of the 2D textures, we use a form of the wavelet phase harmonic correlations of Zhang and Mallat (2021). A phase harmonic of a complex number $z = |z|e^{i\varphi(z)}$ is defined as

$$[z]^k := |z|e^{ik\varphi(z)}.$$

Where the phase was shifted by α in c., it is here scaled by $k \in \{0, 1, 2\}$. As the name suggests, wavelet phase harmonic correlations are correlations on the phase harmonics of a wavelet transform. As with the scattering spectra d., this is a very high-dimensional representation, and by analyzing their properties, informative elements can be selected for a more compact representation. In the experiments we use Model D of (Zhang and Mallat, 2021, Section 5.2), with code from `https://github.com/sixin-zh/kymatio_wph` updated to run in Python 3.10 with PyTorch 2.1.

E Appendix – Ising process

A square lattice Ising process is a binary spin model with a macrocanonical distribution having the energy function $\Phi(x) = -\sum_{(ij)} x_i x_j$ where the sum is over all neighboring pairs (ij) counted once. The boundaries are



(a) True (b) Wav. REG (c) Scat. REG

(d) Gaussian (e) Wav. MF (f) Scat. MF

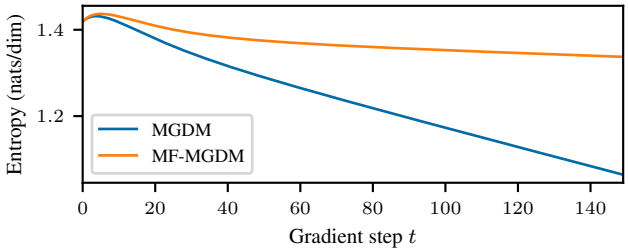
Figure 10: A 256×256 Ising process at temperature 2.4 (a). The MGDM samples were generated using wavelet ℓ^1 norms ((b) and (e)) and scattering moments ((c) and (f)) coefficients. Samples from the regular MGDM (REG) and MF–MGDM (MF) are virtually indistinguishable. For reference, a Gaussian rounded to ± 1 is included (d). All figures except (a) share the same seed.

taken to be periodic, resulting in a torus. Around the critical temperature 2.27, spins tend to cluster, resulting in crystallizing structures as in Figure 10a.

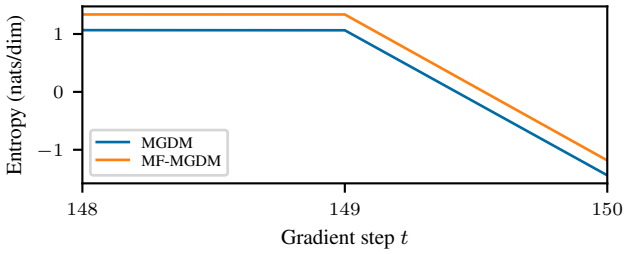
Following Bruna and Mallat (2019), we use wavelet and scattering moments to encode correlations, together with a penalty loss to force the values to ± 1, when sampling by the gradient descent procedure. For MF–MGDM, the batch size is $N = 128$, and the penalty loss is included for each sample and not over the average over the batch. We found that running the descent until the values are sufficiently close to ± 1 has an excessive cost in terms of entropy for both methods, and for certain choices of penalty weights the procedure tends to become stuck in poor local minima. We instead add an element-wise soft rounding $\operatorname{sgn}(x)|x|^{0.05}$ after fewer steps which better preserves entropy and ensures that constraints are close to being satisfied, albeit resulting in a worse energy fit. This, however, does not seem to have an effect on the visual quality of the generated samples.

In Figure 10 we compare samples of resolution 256×256 generated from the two MGDM models for different choices of energy functions. Samples generated from the same noise are visually almost identical.

As with the bubble texture, the models are also compared by computing their entropy on a 32×32 domain (necessary in order to fit the necessary Jacobians in memory). As shown in Figure 11a, MF–MGDM maintains more of its initial entropy. In the final rounding step, a significant loss of entropy occurs (Figure 11b). This is expected, as we are computing the differential entropy of a continuous distribution which approximates a discrete one. Note that the drop is roughly equal for the models, and that MF–MGDM still has a clear advantage in absolute terms. This, together with the observation that generated samples are of the same visual quality, speaks in favor of using the mean-field loss also in the case of approximating high-dimensional discrete distributions.



(a)



(b)

Figure 11: Differential entropies for MGDM approximations of a 32×32 Ising process at temperature 4.0. The graphs have been split up into two for legibility: top figure shows how the entropy evolves up util the soft rounding; bottom shows that the difference in entropy persists during the soft rounding. The energy function used is the average ℓ^1 norm of the wavelet transform.

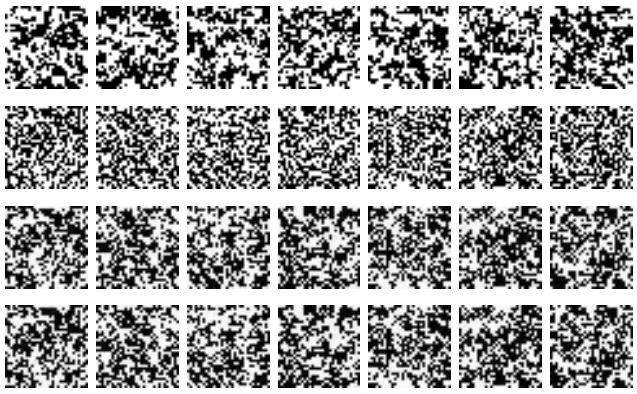


Figure 12: Realizations and approximations of 32×32 Ising process at temperature 4.0. From top to bottom: True, rounded Gaussian, MGDM, MF–MGDM. The energy function for the MGDM models is the average of the ℓ^1 norm of the Wavelet transform, as in Fig. 11. The three last rows share seed column wise.