QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXXVII, NUMBER 2
JUNE 2019, PAGES 269–321
https://doi.org/10.1090/qam/1532
Article electronically published on January 25, 2019

BUILDING A TELESCOPE TO LOOK INTO HIGH-DIMENSIONAL IMAGE SPACES

By

MITCH HILL (*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095*),

ERIK NIJKAMP (*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095*),

AND

SONG-CHUN ZHU (*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095*)

Abstract. In Grenander's work, an image pattern is represented by a probability distribution whose density is concentrated on different low-dimensional subspaces in the high-dimensional image space. Such probability densities have an astronomical number of local modes corresponding to typical pattern appearances. Related groups of modes can join to form macroscopic image basins (known as Hopfield memories in the neural network community) that represent pattern concepts. Grenander pioneered the practice of approximating an unknown image density with a Gibbs density. Recent works continue this paradigm and use neural networks that capture high-order image statistics to learn Gibbs models capable of synthesizing realistic images of many patterns. However, characterizing a learned probability density to uncover the Hopfield memories of the model, encoded by the structure of the local modes, remains an open challenge. In this work, we present novel computational experiments that map and visualize the local mode structure of Gibbs densities. Efficient mapping requires identifying the global basins without enumerating the countless modes. Inspired by Grenander's jump-diffusion method, we propose a new MCMC tool called Attraction-Diffusion (AD) that can capture the macroscopic structure of highly non-convex densities by measuring *metastability* of local modes. AD involves altering the target density with a *magnetization* potential penalizing distance from a known mode and running an MCMC sample of the altered

Received February 17, 2018, and, in revised form, October 18, 2018.
2010 *Mathematics Subject Classification.* Primary 65C40.
This work was supported by DARPA #W911NF-16-1-0579. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ASC170063.
Email address: mkhill@ucla.edu
Email address: enijkamp@ucla.edu
Email address: sczhu@stat.ucla.edu

density to measure the stability of the initial chain state. Using a low-dimensional generator network to facilitate exploration, we map image spaces with up to 12,288 dimensions (64×64 pixels in RGB). Our work shows: (1) AD can efficiently map highly non-convex probability densities, (2) metastable regions of pattern probability densities contain coherent groups of images, and (3) the perceptibility of differences between training images influences the metastability of image basins.

1. Introduction.

1.1. *Motivation.* Representing image patterns requires reconciling the common structure present among images with the variability that exists between images, and addressing this tension is the central theme of Ulf Grenander's pioneering body of work on Pattern Theory [15–18]. As a concrete example, a digit can be written in many different ways, but humans can still recognize a common concept across the change in appearance. Grenander studied classical mathematics early in his career, but came to believe that the models of the time were too rigid to capture the rich variation found in real-world phenomena. Shifting his focus, he initiated the study of Pattern Theory in the 1960s, at a time when almost no literature or known uses existed. His countless contributions have led the field to the prominent role it plays today in many academic disciplines and practical applications.

Stochastic models, where an image I is treated as a sample from a probability density f over the image space, are well-suited for accommodating the tension between structure and variation that exists in real-world patterns. The statistical concept of a probability density f and the physical concept of a diffusion process on the potential energy manifold $V = -\log f$ are equivalent and throughout the paper we use both perspectives interchangeably, although we focus more on the second view. Since I is a random sample from f, image appearance can vary stochastically, but the probability of observing an image is virtually zero except for a small region around the modes of f, enforcing structure in the sampled images. Stochastic image models in high-dimensional spaces are the principle objects of study in Grenander's work.

When modeling image patterns, the true density f is unknown. Grenander realized early in his career that designing an analytical formulation of f from first principles was a hopeless task for real-world patterns. Instead, Grenander sought a family of probability models $\mathcal{P}$ flexible enough to approximate many different pattern densities. Real images, treated as independent samples from f, are used to find a model $p \in \mathcal{P}$ that is a good approximation for f, usually by MLE. Grenander was particularly interested in the family of Gibbs distributions defined on a graph over the pixel lattice, and he validates the capabilities of this family in many experiments. Recent advances have further increased the representational capacity of Gibbs image models (see Section 2.1).

In this paper, we investigate the structure of a learned Gibbs density p (or equivalently, energy $U = -\log p$) trained to model an unknown image density f. During training, the density learns to form modes around the samples of f, and local minima of U can be interpreted as "memories" of the training data, as in Hopfield's model [22]. Regions of the image space separated only by low barriers in U represent groups of images/memories that are conceptually similar. One can imagine the image space as a vast and mostly

empty universe, U as gravitational potential energy, and the local minima of U as dense stars that lie on the pattern manifold. Groups of related local minima separated by low energy barriers (such as different images of the same digit) form connected clusters of pattern images, which are "galaxies" in the image universe (see Figure 1).

FIG. 1. Analogies for energy basins of images in different entropy regimes. Low-entropy images have distinct appearances and create galaxies with macroscopic substructure, like the arms of the spiral galaxy on the left. High-entropy images such as textures cannot be easily distinguished and form wide energy basins with little substructure, like the nebula on the right. See Section 1.2.

Following the approach of Bovier [4], one can formally characterize image galaxies by dividing the image space into *metastable* regions, such that a diffusion process on U mixes over short time-scales within a region, while mixing occurs over long time-scales between regions. In other words, a local MCMC sample of p initiated from an image galaxy will travel in the same galaxy for a very long time, because random fluctuations are enough to overcome small energy barriers within the galaxy, while much larger energy barriers restrict movement between galaxies. This view is closely related to Grenander's jump-diffusion method [16], which uses a combination of local diffusion in a limited region of the state space and global proposals that jump between separate regions of the state space to facilitate sampling. Our primary goal in this paper is to computationally identify metastable regions in an image density while only visiting a few of the local modes within each region, because exhaustive enumeration of modes is computationally infeasible.

The galaxies represent different concepts in the image pattern, and by finding the galaxies of an image density we can reduce the vast high-dimensional image space to a few groups that summarize the major pattern appearances. We are also interested in measuring the energy barriers between galaxies, because they encode similarity between groups. The structure of a learned image density encodes memories of the pattern manifold, but this information is hidden in p and must be recovered through mapping. Landscape structure varies according to the pattern being modeled, as in Figures 1 and 3. In particular, we conjecture that the depth/stability of image basins is related to the human ability to distinguish between pattern images. This idea can be understood by examining the energy landscape of the same pattern at different scales (see Section 1.2).

The formulation, tools, and goals of this paper can all be traced back to Grenander's legacy of research on image models. Grenander was among the first to understand the importance of Gibbs distributions as a flexible and powerful family of models for representing complex data, and he used Gibbs models extensively throughout his work. He is

one of the pioneers of MCMC computing, and his celebrated jump-diffusion method is closely linked with metastable descriptions of energy landscapes. Nonetheless, Grenander faced several major obstacles which prevented him from realizing the full potential of probabilistic image representations. The challenges are listed as follows:

(1) difficulty of defining meaningful potential functions for Gibbs models,
(2) patterns of different scales require separate representations,
(3) sampling from high-dimensional image distributions is expensive,
(4) energy functions of images have highly non-convex structure.

Recent advances in image modeling have made great progress towards resolving the first two issues (see Section 2.1), and this paper tackles the last two difficulties. By overcoming central challenges of Grenander's time, our work is the first to computationally map the structure of Hopfield memories of a Gibbs image distribution. We make several major contributions to the study of probabilistic image models and non-convex energy functions, including:

(1) an MCMC tool for detecting metastable regions of highly non-convex energy landscapes,
(2) a new procedure for mapping the macroscopic structure of non-convex energy landscapes at different resolutions,
(3) a new method for finding low-energy interpolations between local minima in both discrete and continuous energy landscapes,
(4) use of a low-dimensional generator network to facilitate sampling and mapping in the high-dimensional image space,
(5) novel energy-based mappings of pattern concepts in both the image space and the latent space of a generator network,
(6) experimental evidence linking the perceptibility of difference among pattern images and the stability of image basins in a learned landscape.

The paper is organized as follows. In Section 1, we give an overview of our motivation, method, and results. Section 2 summarizes previous work that is relevant to our research. Section 3 introduces Attraction-Diffusion, our proposed MCMC technique, and Section 4 describes a framework for mapping the energy landscape using Attraction-Diffusion. In Section 5, we apply our new method to map the local minima structure of the SK spin-glass Hamiltonian and energy-based image models learned by neural networks.

1.2. *Information scaling and the energy landscape.* Image scale should have a strong influence on the structure of image memories. In one of the central paradigms of pattern representation, Julesz identifies two major regimes of image scale: texture and texton. *Textures* are high-entropy patterns defined as groups of images sharing the same statistics among nearby pixels [23]. *Textons*, on the other hand, are low-entropy patterns, and can be understood as the atomic building elements or local, conspicuous features such as bars, blobs, or corners [24].

As illustrated in Figure 2, texton-scale images have explicit structure that is easily recognizable, and this structure allows humans to reliably sort texton images into coherent groups. Texture-scale images have implicit structure, and it is usually difficult or impossible to find groups among images of the same texture, because no distinguishing features can be identified within a texture ensemble. As image scale increases, the

FIG. 2. Ivy leaves at different scales. As image scale increases from left to right, an increasing variety of image groups can be identified, until one reaches the threshold of perceptibility, after which it becomes difficult to distinguish between images. The fourth scale is close to the perceptibility threshold of humans, while the fifth scale is beyond human perceptibility. A regime transition from explicit, sparse structure to implicit, dense structure occurs as the threshold is crossed. A similar transition occurs in the energy landscape (see Figure 3).

number of recognizable image groups tends to increase until one reaches the threshold of perceptibility, where texton-scale images transition into texture-scale images and humans begin to lose the ability to identify distinguishing features [47]. Beyond the threshold of perceptibility, texture images cannot be told apart or reliably sorted into groups. Change of image scale causes a change in the statistical properties of an image, and we call this phenomenon *Information Scaling*.

We conjecture that Information Scaling is reflected in the structure of the image landscape, and that there is a connection between the perceptibility of differences between pattern images and the stability/depth of local minima images. When the landscape models texton-scale images, where groups among the images can easily be distinguished, we expect to find many separate, stable basins in the landscape encoding the separate appearances of the groups. Landscapes that model texture-scale images, on the other hand, should exhibit behavior similar to human perception and form a single macroscopic basin of attraction with many shallow local minima to encode the texture. By mapping images from the same pattern at multiple scales, we show that the transition in perceptibility that occurs between scales results in a transition in the landscape structure of image memories (see Figure 3).

1.3. *Overview of method and experiments.* Characterizing the structure of energy functions of complex systems in terms of their local minima and the barriers between minima is an important but difficult task that can shed light on the behavior and properties of the system in question. In virtually all cases of interest, the size of the system is so vast that it is impossible to map the landscape by simply evaluating the energy of all possible states. Computational methods are needed to identify local minima and barriers while visiting only a tiny fraction of the system states. We refer to the task of computationally identifying the local minima structure of non-convex energy functions as *Energy Landscape Mapping* (ELM).

Often, the number of local minima is also too vast for full enumeration. On the other hand, macroscopic structures (such as image "galaxies") exist in many non-convex landscapes, even if the local structure is very noisy. Our work proposes a new MCMC "telescope" that can efficiently discover macroscopic structures of complex landscapes in

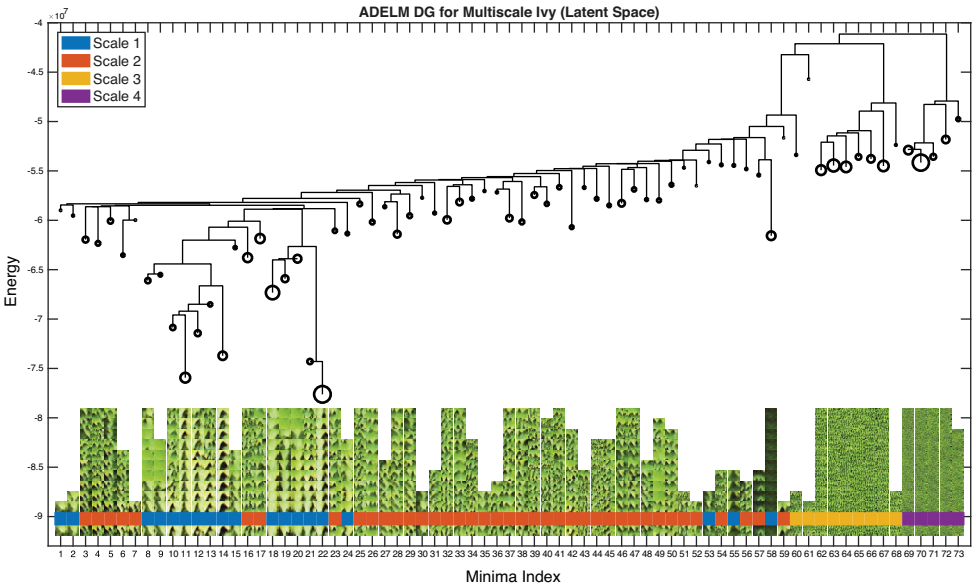


FIG. 3. Landscape of ivy image patches at four different scales. Images from Scale 1 and Scale 2 are textons, while images from Scale 3 and Scale 4 are textures. The texton-scale images account for the majority of the basins in the landscape. More basins are identified for Scale 2 than Scale 1 because Scale 2 has a richer variety of distinct appearances, while the Scale 1 minima have lower energy, since appearances from this scale are more reliable. The texture-scale images form separate basins with little substructure. Basin members from each scale are shown in Figure 4. See Section 5.4 for a full explanation.

both continuous and discrete spaces. This is accomplished by updating MCMC samples using the energy function

$$U_{T,\alpha,X^*}(X) = U(X)/T + \alpha||X - X^*||_2, \tag{1.1}$$

where U is the target energy function, $T > 0$ is temperature, X^* is a known local minimum, and $\alpha > 0$ is the strength of the penalty term. Our method can be viewed as a way of measuring the *metastability* of a local minima in the target landscape U. Metastable basins can be found by carefully tuning α to accelerate the mixing time within basins while still respecting the long time-scales between basins. U_{T,α,X^*} can also be used to find low-energy interpolations between local minima. See Section 3.

In our experiments in Section 5, we map the structure of the DeepFRAME energy function (2.4) and the Co-Op Net energy function (2.13) after the neural network weights have been learned (see Section 2.1 for model descriptions). The DeepFRAME model and Co-Op Net model are good test settings for our mapping algorithm for several reasons.

(1) The energy functions should have multimodal macroscopic structure if the training data can be grouped into different types of images (for example, handwritten digits). These modes can be interpreted as Hopfield associative memories

Multiscale Ivy (Latent Space)

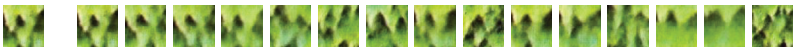
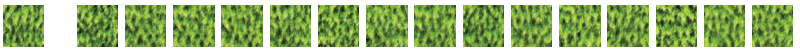
Min. Index	Basin Rep.	Randomly Selected Members (arranged from low to high energy)	Member Count
11			102
22			295
3			37
38			25
63			154
64			117
70			280
71			36

FIG. 4. Minima of multiscale ivy in latent space for the DG depicted in Figure 3. The appearance of randomly selected members is consistent with the appearance of the basin representative.

[22, 48, 49]. The global energy basins will be noisy because of variation possible within the image groups. Our method is designed for this situation.

(2) Since we are mapping energy functions defined over images, the ELM results should roughly correspond to human visual intuition if the mapping is successful. In this case, we can subjectively evaluate our ELM results.

(3) Mapping the local minima structure of DeepFRAME and Co-Op energy functions is a novel application. Much work has been devoted to modeling real data using ConvNet functions, but less work has been done to investigate the structure of these functions after training.

(4) Application to neural network models shows that our method can be successfully used on complex and modern energy functions.

We map image models trained to capture different patterns, and discover a variety of landscape structures. Despite the astronomical number of different image minima, we show that image memories form large structured basins, and that image appearance within global basins is very consistent. Opening up the black box of generative neural networks reveals that the models learn a handful of major image concepts, and our results support the conjecture that perceptibility in the training data influences the stability of image memories in the learned landscape.

2. Related work.

2.1. *Probability models of image patterns.* In practice, the true image density f is unknown, and only training images, which are treated as independent samples from f, are available. To model the image space, one must approximate f by selecting the density

p^* that is "closest" to f from a family of known densities $\mathcal{P}$. When closeness is measured by KL-divergence, this can be accomplished by Maximum Likelihood estimation (see Section 2.2). To obtain an accurate approximation of f, the family of densities $\mathcal{P}$ must be flexible enough to accommodate the variation in the training data.

Gibbs distributions defined on a pixel graph have been widely used as an effective family for modeling patterns of real images [13, 17, 18, 52, 53]. This family of densities has the form

$$p(I) = \frac{1}{Z} \exp\left\{ -\sum_{C \in \mathcal{C}} \varphi_C(I_C) \right\}, \tag{2.1}$$

where $\mathcal{C}$ is the set of cliques of a graph G over the pixel lattice, φ_C are clique potentials over the pixels in clique C, and Z is the normalizing constant. A clique is a group of pixels in which all pairs of pixels are adjacent on G. In early Gibbs image models, the cliques are groups of neighboring pixels and the potentials capture simple clique features, such as consistency of pixel intensity. However, these simple, hand-designed potentials are not capable of synthesizing realistic image patterns. The density (2.1) is very flexible, but the model is useless without a principled way to define clique potentials φ_C that capture relevant features of the target density f.

Zhu et al. address this problem in the FRAME model [53] by using convolutional filters to define clique potentials. The FRAME density has the form

$$p(I) = \frac{1}{Z} \exp\left\{ -\sum_{k=1}^{K} \langle \lambda^{(k)}, H^{(k)}(I) \rangle \right\}, \tag{2.2}$$

where $H^{(k)}(I)$ is a histogram of image responses to convolutional filter k, and $\lambda^{(k)}$ is the potential for filter k. The potential $\lambda^{(k)}$ ensures that the histogram of filter responses $H^{(k)}(I)$ for the sampled image I matches the histogram of filter responses $H^{(k)}(I_{\text{obs}})$ for the training image I_{obs}. The potentials $\lambda^{(k)}$ must be learned. Since filter convolution is a linear projection from the image space to a 1D subspace, matching the sample histogram $H^{(k)}(I)$ to the observed histogram $H^{(k)}(I_{\text{obs}})$ is equivalent to matching the marginal distribution of p to the marginal distribution of f in the 1D subspace of filter k. One can show that $p = f$ if and only if the marginal distribution of p is the same as the marginal distribution of f in all 1D linear subspaces. If the majority of the variation of f is captured by a few marginal directions, matching only these marginal distributions should still give a close approximation for f. The FRAME model learns an image density p by matching the marginal distribution of samples from f in the most relevant filter subspaces.

In the original FRAME model, filters are selected from a pre-defined filter bank, which limits the kinds of patterns that can be represented. There is no guarantee that the filter bank can project onto the most relevant 1D subspaces of f, and synthesis results are poor when filters cannot capture important features of f. Hand-designing filter banks for each new pattern is not a viable solution, because this is just as difficult as hand-designing clique potentials.

Recent trends in the neural network community have shown that learning the filters themselves during training can result in flexible and realistic image models. Including

multiple layers of filter convolution can also lead to significantly better representations of complex data [25, 28]. The DeepFRAME model [31, 48] extends the FRAME model to incorporate these new features. A DeepFRAME density for an image I with D pixels has the form

$$p(I|W) = \frac{1}{Z(W)} \exp\{F(I|W)\}q(I), \tag{2.3}$$

where q is the prior distribution $\mathrm{N}(0, \sigma^2 \mathrm{Id}_D)$ of Gaussian white noise, and the scoring function $F(\cdot|W)$ is defined by a ConvNet with weights W, which must be learned. The normalization constant $Z = \int \exp\{F(I|W)\}\, q(I)dI$ is intractable. The associated energy function has the form

$$U(I|W) = -F(I|W) + \frac{1}{2\sigma^2}||I||_2^2. \tag{2.4}$$

We may interpret $p(I|W)$ as an exponential tilting of q which has the effect of mean shifting. The non-linearity induced by the activation functions between network layers is essential for successful representation of real images.

When the the activation functions are rectified linear units (ReLU), $F(\cdot|W)$ is piecewise linear in I, and the borders between linear regions are governed by the activations in the network [33]. Let $\Omega_{\delta,W} = \{I : \sigma_k(I|W) = \delta_k, 1 \le k \le K\}$, where W gives the network weights, K is the number of activation functions in the entire network, $\sigma_k(I|W) \in \{0, 1\}$ indicates whether activation function k turns on for image I, and $\delta = (\delta_1, \ldots, \delta_K) \in \{0, 1\}^K$. Since $F(I|W)$ is linear on $\Omega_{\delta,W}$ for all δ, the energy can be written as

$$U(I|W) = -(\langle I, B_{\delta,W}\rangle + a_{\delta,W}) + \frac{1}{2\sigma^2}||I||_2^2 \tag{2.5}$$

for some constants $a_{\delta,W}$ and $B_{\delta,W}$, which shows that $I \sim \mathrm{N}(\sigma^2 B_{\delta,W}, \sigma^2 \mathrm{Id}_D)$ on $\Omega_{\delta,W}$ and that $p(I|W)$ is piecewise Gaussian over the image space. This analysis also characterizes the local minima of $U(I|W)$. Let $\mu_{\delta,W} = \sigma^2 B_{\delta,W}$ be the Gaussian mean vector for piece (δ, W). The local modes are then simply $\{\mu_{\delta,W} : \mu_{\delta,W} \in \Omega_{\delta,W}\}$, the Gaussian modes that are contained within their own piece. However, there is no guarantee that the Gaussian piece $\Omega_{\delta,W}$ contains its mode $\mu_{\delta,W}$, and the number of Gaussian pieces is extremely large, so mapping a DeepFRAME model by identifying all Gaussian pieces is not viable.

Early image models often employ different representations to cover the scale spectrum. Sparse basis functions can effectively capture the features of texton images, while MRF distributions are more suitable for representing texture patterns [39, 47, 53]. The DeepFRAME density incorporates aspects of both families, because the filters serve as both implicit features and sparse basis functions for image synthesis [31, 48]. The DeepFRAME model provides a unified way to represent image patterns at many different scales, but we still expect to identify different structures in the image landscape across the scale spectrum.

2.2. *Learning image models with Maximum Likelihood.* Maximum Likelihood is a principled and widely-used method for estimating the parameters of a distribution from an observed sample. The log-likelihood of the DeepFRAME density $p(I|W)$ given i.i.d.

images $\{I_i\}_{i=1}^n$ is

$$l(W) = \frac{1}{n}\sum_{i=1}^{n} \log p(I_i|W) = -\log Z(W) - \frac{1}{n}\sum_{i=1}^{n} U(I_i|W), \tag{2.6}$$

and maximizing $l(W)$ yields the Maximum Likelihood Estimate (MLE) W^* of the model parameters. Observe that

$$KL\left[q(I)\,||\,p(I|W)\right] = \mathrm{E}_q\left[\log \frac{q(I)}{p(I|W)}\right] = \mathrm{E}_q[\log q(I)] - \mathrm{E}_q[\log p(I|W)].$$

The term $\mathrm{E}_q[\log q(I)]$ does not depend on W and the Law of Large Numbers shows that

$$\mathrm{E}_f[\log p(I|W)] \approx \frac{1}{n}\sum_{i=1}^{n} \log p(I_i|W) = l(W).$$

Therefore maximizing the log-likelihood (2.6) to find the MLE is equivalent to finding the value of W that minimizes the KL Divergence between $p(I|W)$ and the true data distribution f.

One can solve for W by maximizing $l(W)$ with gradient ascent. The intractable partition function $Z(W)$ is a major obstacle when evalutating $\nabla l(W)$. Fortunately, the gradient of $\log Z(W)$ can be expressed in closed form:

$$\frac{d}{dW}\log Z(W) = -\,\mathrm{E}_{p(I|W)}\left[\frac{\partial}{\partial W}U(I|W)\right]. \tag{2.7}$$

The expectation is still intractable, but it can be estimated by drawing MCMC samples $\{Y_i\}_{i=1}^m$ from the current distribution $p(I|W)$ and using a Law of Large Numbers approximation. This yields the stochastic gradient

$$\widetilde{\nabla} l(W) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W}U(Y_i|W) - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial W}U(I_i|W) \tag{2.8}$$

which can be used to iteratively solve for W.

Maximum Likelihood learning requires repeatedly drawing MCMC samples from the current distribution $p(I|W)$ to estimate the gradient of the normalizing constant. This can be computationally expensive in the image space, because dimension scales with the square of the image width, so even small images are high-dimensional. Langevin Dynamics are often the sampling method of choice when conducting MCMC in high dimensions. Updating MCMC samples according to the Langevin Equation

$$I_{t+1} = I_t + \frac{\varepsilon^2}{2}\frac{d}{dI}\log p(I_t|W)dt + \varepsilon B_t, \tag{2.9}$$

where $B_t \sim \mathrm{N}(0, \mathrm{Id}_D)$, preserves the distribution of p after a Metropolis-Hastings correction [12,35]. The gradient term in the Langevin Equation leads to faster convergence than methods such as Random-Walk Metropolis-Hastings and Gibbs sampling, which make no use of the local landscape geometry.

Even with Langevin Dynamics, it is infeasible to obtain true independent samples of p each time the model is updated. Contrastive Divergence (CD) [21] and Persistent Contrasitve Divergence (PCD) [41] are two common methods of obtaining approximate samples of p. In CD, the training images are used as the initial states of MCMC samples

from p, while in PCD the images from the previous training iteration are used as the initial states. A sketch of the DeepFRAME training algorithm with PCD updates is given in Algorithm 1.

Algorithm 1: DeepFRAME Learning Algorithm

input : Observed images $\{I_i\}_{i=1}^n$, number of latent samples m, number of Langevin iterations K, step size $\delta > 0$, number of learning steps S, initial weights W_0, initial persistent synthesized images $\{Y_i\}_{i=1}^m$.

output: Weights W^* for energy $U(I|W)$.

for $s = 1 : S$ **do**

1. Using equation (2.9), apply K Langevin updates to the images $\{Y_i\}_{i=1}^m$ with the current energy $U(I|W_{s-1})$.
2. Use a mini-batch $\{I_i\}_{i=1}^m$ of training data and revised images $\{Y_i\}_{i=1}^m$ to update W according to
$$W_s = W_{s-1} + \delta \widetilde{\nabla} l(W_{s-1})$$
where $\widetilde{\nabla} l(W)$ is the log-likelihood gradient in (2.8).

In this paper, we are interested in mapping the local minima structure of a learned energy U of the form (2.4) for a DeepFRAME density (2.3) which is trained to model the true, but unknown, image density f. Unfortunately, mapping a DeepFRAME energy directly is problematic because the energy function learns many accidental low-energy regions that can obscure the relations between image basins. Since training relies on CD or PCD, the DeepFRAME energy only observes warm-start images that are already quite close to the pattern manifold. The landscape structure of a DeepFRAME energy is only meaningful in a small region around the pattern manifold and vast low-energy basins can form in remote regions of the image space. Using a Gibbs sampler instead of Langevin Dynamics during training can alleviate the problem at the cost of efficiency and scalability.

2.3. *Generator networks, cooperative learning, and a cooperative energy function.* To overcome deficencies found in DeepFRAME energy functions, we introduce a generator network [14, 20] which learns a set of weights to transform a trivial latent distribution into a distribution over the image space that approximates the pattern manifold.

Let the D-dimensional image data I follow the distribution

$$I \sim \mathrm{N}(g(Z|W_2), \tau^2 \mathrm{Id}_D) \tag{2.10}$$

with $Z \sim \mathrm{N}(0, \mathrm{I}_d)$ for $d \ll D$, variance parameter τ^2, and weights W_2 of a ConvNet function g. The joint energy function for (I, Z) has the form

$$U(I, Z|W) = \frac{1}{2\tau^2}||I - g(Z|W)||^2 + \frac{1}{2}||Z||^2$$

which is simply the sum of the Gaussian energy functions of Z and $I|Z, W$. The energy function of the conditional variable $Z|I, W$ is $U_{Z|I,W}(z) = U(z, I|W)$, since the posterior distribution $Z|I$ is proportional to the joint distribution of (I, Z).

The latent factors $\{Z\}_{i=1}^n$ are unknown, and W must be learned by maximizing the observed data log-likelihood, which corresponds to maximizing the function

$$l(W) = \sum_{i=1}^n \log p(I_i|W) = \sum_{i=1}^n \log \int p(I_i, Z|W) dZ$$

that integrates the latent factors out of the joint distribution. This loss cannot be computed directly, but the gradient of the log-likelihood can be rewritten as

$$\frac{\partial}{\partial W}\log p(I|W) = -\mathrm{E}_{Z|I,W}\left[\frac{\partial}{\partial W}U(I,Z|W)\right],$$

so the log-likelihood gradient can be estimated by drawing MCMC samples of $Z|I,W$, the latent factors conditioned on the observed data, using the current weight W. Langevin Dynamics can be used to sample from $Z|X_i,W$, and the Langevin update equation is

$$Z_{t+1} = Z_t + \frac{\varepsilon^2}{2}\left(\frac{1}{\tau^2}(I_i - g(Z_t|W))\frac{\partial}{\partial Z}g(Z_t|w) - Z_t\right) + \varepsilon B_t \tag{2.11}$$

for $B_t \sim \mathrm{N}(0,\mathrm{Id}_d)$ and step size ε for $t = 1,\ldots,K$ iterations. One Z_i is inferred for each observed image I_i. PCD is used during training, so MCMC sampling in each new inference phase is started from the Z_i of the previous inference phase. Once the Z_i have been sampled from $p(Z|I_i,W)$, the weights W can be updated with

$$\widetilde{\nabla} l(W) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}U(I_i,Z_i|W) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tau^2}(I_i - g(Z_i|W))\frac{\partial}{\partial W}g(Z_i|W) \tag{2.12}$$

in the second phase of the algorithm. The inference phase uses a back-propagation gradient $\frac{\partial}{\partial Z}g(Z|W)$, while the learning phase uses a back-propagation gradient $\frac{\partial}{\partial W}g(Z|W)$. The calculations required to obtain $\frac{\partial}{\partial Z}g(Z|W)$ are needed as part of the calculation of $\frac{\partial}{\partial W}g(Z|W)$, so both phases can be implemented in a similar way.

The Co-Op Net Algorithm [49] provides a way to simultaneously learn the weights W_1 of a DeepFRAME energy function $U(\cdot|W_1)$ and the weights W_2 of a generator network $g(\cdot|W_2)$ for any dataset. During training, the generator g learns to mimic the manifold of U, while the energy function U learns to model the real data. Following [49], the DeepFRAME network in the Co-Op Net model is referred to as the *descriptor* network, because the energy function encodes a description of image features, which are expressed through filter activations. The descriptor and generator networks provide a natural warm-start initialization for the MCMC sampling phase of the partner network. A sketch of the Co-Op Net Training is presented in Algorithm 2.

By composing a generator network $g(Z|W_2)$ and descriptor energy $U(I|W_1)$, we can define a new energy function

$$U(Z|W_1,\, W_2) = U(g(Z|W_2)|W_1) \tag{2.13}$$

over the latent space. This formulation is very similar to the DGN-AM model [36] (see Section 2.9). Sampling in the low-dimensional latent space vastly reduces computational cost, providing a way to efficiently explore the pattern manifold of realistically-sized images. We find that Metropolis-Hastings can actually be more efficient than Langevin Dynamics when sampling from a low-dimensional latent space because Metropolis-Hastings only requires a relatively inexpensive forward pass network evaluation while Langevin Dynamics requires a forward and backward pass to compute the gradient.

Algorithm 2: Cooperative Learning Algorithm

input : Observed images $\{I_i\}_{i=1}^n$, number of latent samples m, number of Langevin iterations K, descriptor step size $\delta_1 > 0$, generator step size $\delta_1 > 0$, number of learning steps S, initial weights $W_{1,0}$ for descriptor and $W_{2,0}$ for generator.

output: Weights W_1^* for descriptor energy $U(I|W_1)$ and W_2^* for generator $g(Z|W_2)$.

for $s = 1 : S$ **do**

1. Draw i.i.d. samples $\{Z_i\}_{i=1}^m$ from the latent distribution $\mathrm{N}(0, \mathrm{Id}_d)$ of the generator network $g(Z|W_{2,\,s-1})$. Compute images $\{Y_i\}_{i=1}^m$, where $Y_i = g(Z_i|W_{2,\,s-1})$.
2. Using equation (2.9), apply K Langevin updates to the images $\{Y_i\}_{i=1}^m$ with the current energy $U(I|W_{1,\,s-1})$ to obtain revised images $\{\tilde{Y}_i\}_{i=1}^m$.
3. Using equation (2.11), apply K Langevin updates to the latent factors $\{Z_i\}_{i=1}^m$ with the current weights $W_{2,\,s-1}$, where the revised $\tilde{Y}_i$ from the previous step is the conditional image for each Z_i.
4. Use a mini-batch $\{I_i\}_{i=1}^m$ of training data and revised images $\{\tilde{Y}_i\}_{i=1}^m$ to update W_1 according to
$$W_{1,\,s} = W_{1,\,s-1} + \delta_1 \widetilde{\nabla} l_1(W_{1,\,s-1})$$
where $\widetilde{\nabla} l_1(W)$ is the log-likelihood gradient in (2.8).
5. Use revised latent factors $\{Z_i\}_{i=1}^m$ and *revised* images $\{Y_i\}_{i=1}^m$ to update W_2 according to
$$W_{2,\,s} = W_{2,\,s-1} + \delta_2 \widetilde{\nabla} l_2(W_{2,\,s-1})$$
with $\widetilde{\nabla} l_2(W)$ is the log-likelihood gradient in (2.12).

Interestingly, it appears that the structure of image memories in the energy (2.13) is more meaningful than the structure of memories in (2.4), because concatenating the generator and descriptor networks reduces the number of accidental low-energy regions found between pattern minima in the raw DeepFRAME energy (compare Figures 21 and 22). The energy (2.13) provides a way to characterize both the image space and the latent space of a generator network. Previous works have identified a handful of minima in the latent space using a similar energy function [36], but our work is the first to systematically explore and map the structure of a latent generator space.

2.4. *Macroscopic structure of non-convex landscapes.* Energy functions associated with complex systems are often non-convex, and the degree of non-convexity in the energy landscape varies depending upon the system in question. In some settings, the landscape has only slight non-convexity, and optimizing the non-convex energy function leads to a solution close to the global minimum, as in [30]. In contrast, the loss surfaces of ConvNet classification and regression functions are highly non-convex, because symmetry-breaking occurs early in training as the filters compete to represent different features of the data. Eventually, the filters settle into one of an astronomical number of distinct parameterizations with nearly equivalent loss [8].

Often, a highly non-convex landscape can have simple and recognizable global structure. A well-known example is the "funnel" structure of potential energy surfaces associated with protein folding [37]. A funnel shape is well-suited for guiding an unfolded or partially-folded protein to its native state. Weakly stable intermediate states might occur along the folding path, but random perturbations from the environment are enough to upset these shallow minima and allow the folding process to continue. Once the protein

has reached its native state, its configuration is stable and resistant to small perturbations. The macroscopic landscape has a single global basin, despite the astronomical number of weakly stable intermediate states along the "sides" of the funnel.

If the large-scale structure of an energy landscape is dominated by a manageable number of global basins, it should be possible to identify these energy basins and to estimate the energy barriers between them. In image landscapes, the global funnels represent the different concepts in the image patterns, since related image minima are separated by small energy barriers. Mapping only the large-scale features while ignoring local irregularities in a landscape is a key innovation of our paper. This approach distinguishes our work from previous efforts to characterize non-convex landscapes such as [2, 9, 44, 51], which attempt to identify all local minima (or the N lowest-energy minima) in the landscape, no matter how weak the basin of attraction. By focusing on macroscopic features, we define a new ELM framework that scales well with landscape dimension and/or complexity (see Sections 3 and 4).

2.5. *Minimum energy path estimation.* Energy barriers between local minima can be used to quantify "closeness" of minima in the landscape, because the barriers provide a measure of the geodesic distance along the energy manifold between minima. Euclidean distance in the state space is a very poor approximation of geodesic distance. Wide, noisy basins can contain points that are far apart in Euclidean space, while two points which are nearby in Euclidean space might be separated by a large energy barrier, as in Figure 7.

Finding energy barriers involves approximating the Minimum Energy Pathway (MEP) between the minima. The simplest approach is to find the maximum energy along the linear 1D subspace between two minima [19], but this often significantly overestimates the true energy barrier between points on the manifold, even over short distances (see Figures 11 and 25). The chemical physics community has developed two major families of methods for MEP estimation. One branch of MEP methods, known as *single-ended* methods, involves starting at a known local minimum and finding a transition state between minima by following the path of slowest ascent along the minimum-eigenvalue direction of the local Hessian [5, 40, 50]. This method fails when Hessian information is not available or cannot be accurately approximated.

Another branch of MEP methods, called *double-ended* methods, involves refining a *chain-of-states* ($F_0 = X_a, F_1, \ldots, F_N, F_{N+1} = X_b$) between two minima X_a and X_b by minimizing the objective function

$$L(\{F_j\}_{j=1}^N) = \sum_{j=1}^{N} U(F_j) + \sum_{j=0}^{N} \frac{Nk}{2} ||F_{j+1} - F_j||_2^2, \tag{2.14}$$

where U is the target energy and $k > 0$ is a "spring force" between successive chain states [11, 26]. Double-ended MEP methods require an initialization path, which by default is the 1D linear subspace between minima, since no other choices are available. Optimizing the loss (2.14) leads to misleading paths where the 1D energy barrier between successive images in the chain is significantly higher than the energy of the images in the chain. Modifications such as the Nudged Elastic Band (NEB) and Doubly-Nudged Elastic Band (DNEB) methods [45] have been introduced to improve optimization by

projecting energy and spring gradients onto the perpendicular and parallel components of the current path direction, respectively. NEB and DNEB require numeric gradients and cannot be used in discrete spaces.

MEP methods have been successfully used to map the energy landscape of stable configurations of molecular systems [44]. Similar methods have been applied to machine learning problems [2, 9], but the results yield an overabundance of local minima and trivial, single-basin macroscopic structure. Our approach is related to the double-ended MEP methods, although we do not try to find the MEP explicitly. On the other hand, the barriers estimated by our method are often significantly lower than the barriers estimated by MEP methods (see Figure 25), and our method can be used for MEP estimation in both discrete and continuous spaces. More importantly, we aim to formulate a more natural criterion for evaluating the "closeness" of two minima, based not on raw barrier height but on the stability of local minima under the time-evolution implied by the energy function.

2.6. *Generalized Wang-Landau Algorithm.* Another approach to mapping non-convex energy functions is a version of the Generalized Wang-Landau (GWL) Algorithm [1, 29, 46] which penalizes repeated visits to pre-defined energy bins within the basin of attraction of a local minimum. An MCMC sample is updated using the time-inhomogeneous, modified Metropolis-Hastings acceptance probability

$$\alpha(S \to S^*) = \min\left(1, \frac{Q(S^* \to S)P(S^*)}{Q(S \to S^*)P(S)} \exp\left\{\gamma(N_{\varphi(S)} - N_{\varphi(S^*)})\right\}\right), \tag{2.15}$$

where P is the target density, Q is the transition probability, $\varphi(S)$ gives the indices (i, j) of the basin i and energy spectrum j to which S belongs, $N_{(i,j)}$ is the number of previous visits to bin (i, j), and $\gamma > 0$ is a penalty for repeated visits to the same bin.

In theory, this algorithm should result in a stationary distribution that visits each energy bin within each basin of attraction in the landscape with equal probability. Barriers between minima can be estimated by locating and refining transition states along the MCMC path. Zhou [51] demonstrated that the GWL Algorithm can be effective in moderately-sized discrete landscapes by mapping the local minima structure of 100-dimensional SK spin-glasses. The GWL method has also been applied successfully to small-scale machine learning problems [38]. However, the GWL Algorithm is ineffective in complex landscapes where the number of distinct local minima is too large for a full enumeration. Our new MCMC method addresses this problem by grouping minima that are separated only by small barriers, which greatly reduces the complexity of the landscape. The GWL Algorithm and our ELM method can be used together, although in the experiments presented in this paper, the GWL penalty was not necessary.

2.7. *Disconnectivity Graphs.* After a mapping is completed, it is useful to visually summarize the local minima and barriers that have been discovered. Visualizing all barriers between minima in a meaningful way is often an impossible task, because it is difficult to concisely represent the complex pairwise relations between the minima. *Disconnectivity Graphs* [3], or DG's, are a widely-used tool for displaying the most important features of an energy landscape. DG's reduce the complexity of the visualization

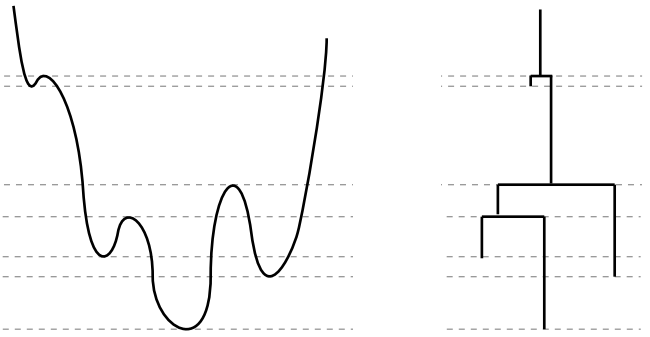


FIG. 5. Illustration of Disconnectivity Graph construction. A 1D energy landscape (left) and its associated DG (right), which encodes minima depth and the lowest known barrier between basins.

task by displaying only the *lowest* barrier at which two groups of minima merge in the landscape.

The leaf nodes in the DG represent local minima in the landscape, and the non-leaf nodes are placed at the lowest-energy barrier at which the basins of the child nodes merge (see Figure 5). Each child node has a single parent node, and the entire DG has a tree structure. The non-leaf nodes are often interpreted as "superbasins" [3] of attraction which are composed of basins of attraction with similar properties. The main focus of our work is to identify superbasins of attraction without identifying all of the local minima within the superbasin.

Figure 6 shows a 2D landscape visualizing the loss of a Gaussian Mixture Model (GMM) as all but two mean parameters are held fixed. In this case, it is easy to see how the structure of the DG reflects the structure of the landscape, since we can visualize the loss function directly. In virtually all real cases, the landscape cannot be directly visualized or exhaustively explored via grid search, but high-dimensional landscape features can still be displayed effectively with a DG.

A major issue with the DG visualization is the greedy nature of the branch-merging step. Merging basins at the lowest possible energy can prevent the appearance of true landscape features in the DG, because lower-energy groups of minima tend to disrupt the structure among higher-energy groups of minima. Nonetheless, DG's are a simple and often effective way of displaying the shape and connectivity of a landscape.

2.8. *Landscape magnetization.* Chaudhari and Soatto [7] use t-SNE [42] to visualize the behavior of the energy function

$$U^*(X) = U(X) + h^\intercal X \tag{2.16}$$

of a spin-glass Hamiltonian U subject to a random magnetization force $h^\intercal X$. As $||h||_2$ increases, the local minima structure of the magnetized landscape goes from a phase where the number of distinct minima is too large for enumeration, through a phase where a manageable number of macroscopic features emerge, to a final phase of trivialization

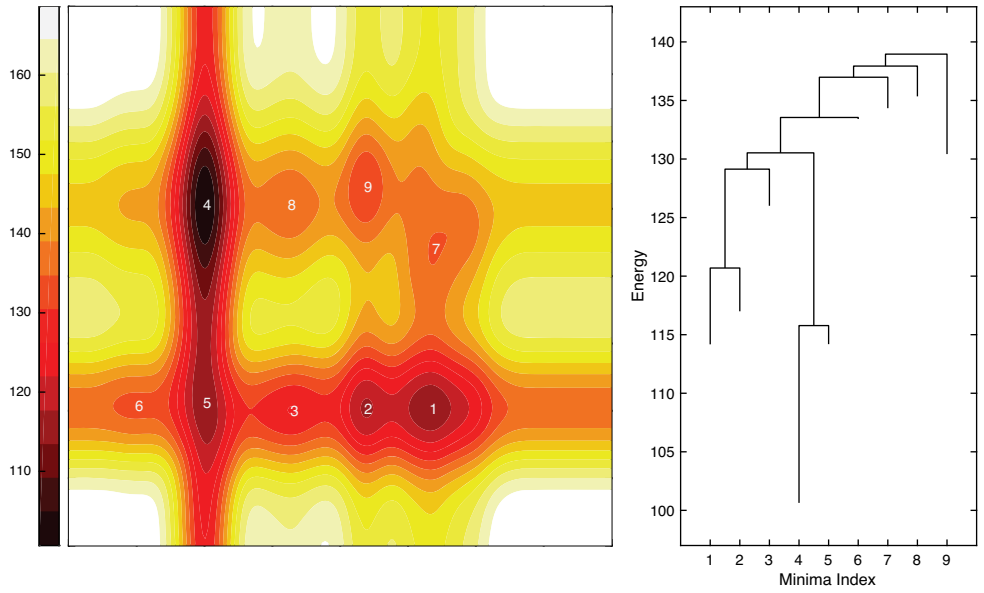


FIG. 6. Landscape Visualization (left) and DG (right) of 2D landscape for GMM mean parameters.

where all minima merge into a single basin. The same behavior occurs during our mapping procedure as α is increased in the altered energy (1.1). The random magnetization h can be interpreted as a version of our penalty which uses a random distribution over the magnetization force α and target state X^*, because the Langevin Equation

$$dX(t) = -\left(\nabla U(X(t)) + h\right) dt + \sqrt{2}\, dB(t) \tag{2.17}$$

associated with (2.16) has the same dynamics as our energy (1.1) when $\alpha = ||h||_2$ and $X^* = X + ch$ for any scalar $c \neq 0$. The authors only characterize the energy landscape using t-SNE plots, and do not attempt to systematically find basins of attraction and barriers in the landscape.

Chaudhari et al. [6] present a modification of the energy function that is similar to our modification to improve training for neural networks. The authors use an altered distribution

$$P_{\gamma,X^*}(X) = \frac{1}{Z_{\gamma,X^*}} \exp\left\{-\left(U(X) + \gamma||X - X^*||_2^2\right)\right\}, \tag{2.18}$$

where X^* is the *current* location and $\gamma > 0$ is a regularization penalty, to find an entropy-biased gradient which favors movement toward wide, flat valleys in the landscape of U. As in our energy (1.1), the penalty term is used to overcome local irregularities, but the interpretations and applications are very different. In [6], the altered density (2.18) is used for the conventional purpose of training of network parameters, while we use the altered energy (1.1) as a metric for metastability and as a tool for mapping landscape structure.

2.9. *Activation Maximization.* Our experiments on image models are closely related to the Activation Maximization (AM) field of neural network research [10, 32, 34]. AM

applications search for images that maximize the response of a neuron or channel in a trained network, which is equivalent to searching for local modes in the Gibbs distribution defined by neuron response. In particular, the model which we focus on in Section 5 is nearly identical to the DGN-AM model [36], where a generator neural network is used to facilitate exploration of a complex neural network energy function. We learn our generator and energy network jointly using the method of Xie et al. [49], while the DGN-AM model uses separate, pre-trained generator and energy networks.

Our work differs from the AM literature in several important ways. Previous AM works only identify a handful of local minima in the energy landscape, and do not attempt to systematically identify the basins of attraction and the structure among these basins. We show not only that neural networks can learn realistic image memories, but also that structure of image memories in the energy landscape reflects human visual intuition. AM applications generally use neurons from pre-trained classifier neural networks as the energy function, while we train our networks specifically to learn an energy function (and generator network) for a training dataset of our choice.

3. Attraction-Diffusion.

3.1. *Introduction to Attraction-Diffusion.* We propose a new method for characterizing the relative stability of local minima of an energy function, which we call *Attraction-Diffusion* (AD). Given an energy function U and two local minima, one minima is designated as the starting location X_0 and the other as the target location X^*. An MCMC sample is initiated from X_0 using an altered density

$$p_{T,\alpha,X^*}(X) = \frac{1}{Z_{T,\alpha,X^*}} \exp\{-(U(X)/T + \alpha||X - X^*||_2)\} \tag{3.1}$$

whose energy function is the sum of the original energy U and a "magnetization" term penalizing the distance between the current state and the target location. T gives the temperature of the system, while α is the strength of the "magnetic field" penalizing distance from the target minimum. The roles of starting and target location are arbitrary and diffusion in both directions is possible. The space of X can be continuous or discrete.

By adjusting the value of α and T, the altered landscape can be tuned so that a diffusion path can overcome small obstacles in the original landscape while remaining trapped in strong basins. If the Markov chain comes within a close distance of the target state, then the starting state belongs to the same energy basin as the target state at an energy resolution implicitly defined by the strength of magnetization. If the chain cannot improve on the minimum distance between the previous states of the chain and the target state for M consecutive iterations, then there must be an energy barrier between the starting and target location that is stronger than the force of the magnetization. Figure 7 demonstrates the basic principles of AD in a simple 1D landscape with two global basins.

AD can also be used to estimate the MEP and the energy barrier between minima, since the maximum energy along a successful diffusion path is an upper bound for the minimum barrier height. This estimate can be refined by setting α just above the threshold where the diffusion path fails to reach the target. By using a *local* MCMC method such as

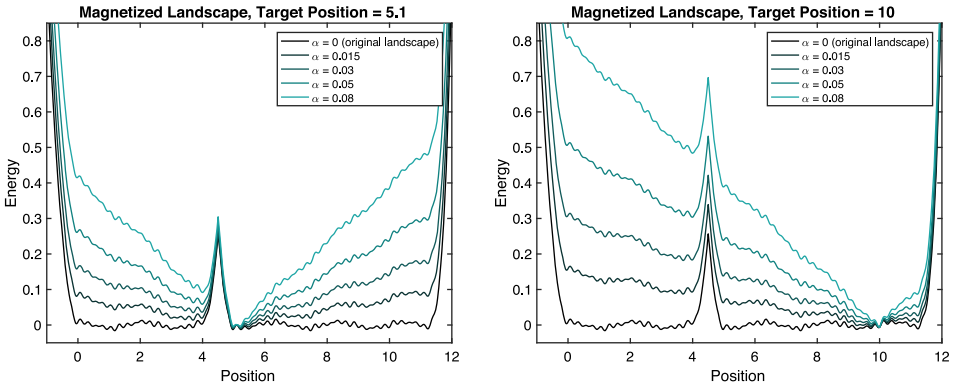


FIG. 7. Magnetization of a toy 1D landscape with target positions $X = 5.1$ (left) and $X = 10$ (right). The original landscape has two flat and noisy basins. Both target positions belong to the same basin, even though they are distant in Euclidean space. The magnetized landscapes have easily identifiable minima, and preserve the large barrier separating the two basins. Since diffusion in the left-hand landscape initiated from $X = 10$ will reach $X = 5.1$, and vice versa in the right-hand landscape, these points belong to the same basin. Low-temperature diffusion initiated from the left of the barrier will be unable to reach the target position in either landscape.

Random-Walk Metropolis-Hastings, Component-Wise Metropolis Hastings, Gibbs sampling, or Hamiltonian Monte Carlo [35], one can limit the maximum Euclidean distance between points in the diffusion path and ensure that the step size is small enough so that the 1D landscape between successive images is well-behaved. An AD chain moves according to geodesic distance in the magnetized landscape, which should be similar to geodesic distance in the raw landscape as long as the strength of magnetization is not too strong.

The choice of the L_2-norm as the magnetization penalty is motivated by the observation that $\frac{d}{dX}||X||_2 = X/||X||_2$, which means that the AD magnetization force points towards the target minimum with uniform strength α throughout the energy landscape. This can be seen in the Langevin Equation

$$dX(t) = -\left(\nabla U(X(t))/T + \alpha \frac{X(t) - X^*}{||X(t) - X^*||_2}\right) dt + \sqrt{2}\, dB(t) \tag{3.2}$$

associated with the magnetized dynamics. An L_1 penalty would probably give similar results. The penalty $\alpha||X - X^*||_2^2$ would *not* have desirable properties because the strength of magnetization would depend on the distance between the points, and the magnitude of alteration would vary throughout the landscape.

The magnetization term in (3.1) is similar to the spring term from the chain-of-states objective (2.14), except that our magnetization force is always pointing to the target minimum X^* with uniform strength α, while the spring force points to the next image in the chain and gets stronger when the distance between images increases. Despite the similarity in the energy functions, AD is most naturally formulated *not* as a way

of estimating the MEP between minima, but as a way of detecting *metastability* (see Sections 3.2 and 3.3) in an energy landscape.

3.2. *Magnetization of the Ising model.* The AD penalty term is closely related to the magnetization term found in energy functions from statistical physics. Consider the N-state magnetized Ising energy function

$$U_{T,H}(\sigma) = -\frac{1}{T}\sum_{(i,j)\in\mathcal{N}} \sigma_i\sigma_j - H\sum_{i=1}^{N}\sigma_i, \tag{3.3}$$

where $\sigma_i = \pm 1$, $\mathcal{N}$ is the set of neighboring nodes, $T > 0$ gives the temperature, and H gives the strength of an external magnetic field. This energy function is sometimes parameterized by the slightly different form $U_{T,H}(\sigma) = \frac{1}{T}\left(-\sum \sigma_i\sigma_j - H\sum\sigma_i\right)$, but the same properties and diagrams hold either way. The first term $-\frac{1}{T}\sum\sigma_i\sigma_j$ is the energy function of the standard Ising model, and $-H\sum\sigma_i$ represents a uniform magnetic field with strength H acting on each node. When $H > 0$, the field has a positive magnetization, encouraging every node to be in state $+1$. In this case, $U_{T,H}$ can be rewritten as

$$\begin{aligned} U^*_{T,H}(\sigma) &= U_{T,H}(\sigma) + NH \\ &= -\frac{1}{T}\sum_{(i,j)\in\mathcal{N}}\sigma_i\sigma_j + H\sum_{i=1}^{N}(1-\sigma_i) \\ &= -\frac{1}{T}\sum_{(i,j)\in\mathcal{N}}\sigma_i\sigma_j + H||\sigma - \sigma^+||_1, \end{aligned}$$

where σ^+ is the state with $\sigma_i^+ = 1$ for all nodes. The probability distribution defined by $U^*_{T,H}$ is the same as the distribution defined by $U_{T,H}$ because they differ only by a constant. Similarly, when $H < 0$ and the magnetic field is negative, the energy function can be rewritten as

$$U^*_{T,H}(\sigma) = -\frac{1}{T}\sum_{(i,j)\in\mathcal{N}}\sigma_i\sigma_j + |H|\,||\sigma - \sigma^-||_1,$$

where σ^- is the state with all $\sigma_i^- = -1$. This shows that the role of H in the magnetized Ising model is the same as the role of α in (3.1), because $U^*_{T,H}$ is the sum of the unmagnetized Ising energy and a term that penalizes distance to either σ^+ or σ^-, the mirror global minima. Introducing the magnetization term upsets the symmetry of the standard Ising energy function and causes either σ^+ or σ^- to become the sole global minimum, depending on the sign of H.

The behavior of the system with respect to the parameters (T, H) can be represented by the simple phase diagram in Figure 8. The dot is the critical temperature of the system, and the solid line is a first-order phase transition boundary. When the parameters of the system are swept across the first-order transition boundary, a discontinuous change in the state space occurs as the system flips from a predominantly positive state to a predominantly negative state, or vice versa. On the other hand, sweeping the magnetic field H across 0 above the critical temperature results in a smooth transition where positive and negative nodes coexist [27].

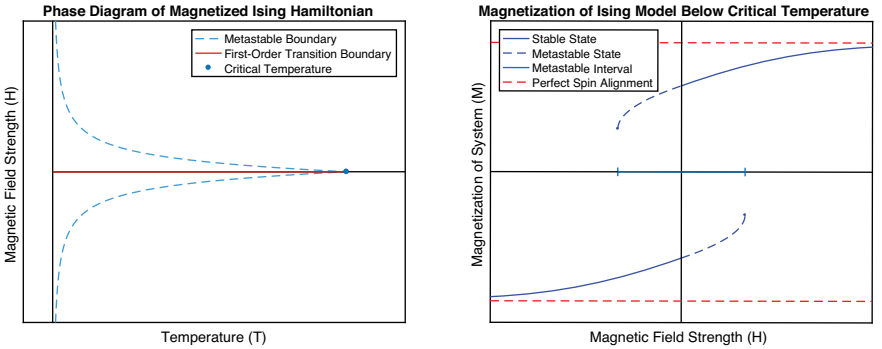


FIG. 8. Left: Phase diagram of the magnetized Ising model. Below the critical temperature, sweeping the magnetic field H from positive to negative (or vice versa) results in a jump between the basins of σ^+ and σ^-. However, if the magnetization force is weak, states in the opposite basin can remain stable for long time periods. Right: Magnetization $M = \sum_i \sigma_i$ as a function of H for a fixed $T^* < T_c$. The metastable interval is the region between the dashed lines along the vertical line $T = T^*$ in the left figure.

Let $H > 0$ be a weak magnetic field, and suppose the temperature T is below the critical temperature T_c. In this situation, a phenomenon known as metastability can occur. If the system is initialized from a random configuration (each node $+1$ or -1 with probability $1/2$), the influence of the magnetic field will cause the system to collapse to σ^+, or a nearby predominantly positive region of the state space, with high probability. However, if the system is initialized from σ^-, and if H is sufficiently small, the system will exhibit metastability, because magnetic force H will be unable to overcome the strength of the bonds in σ^-, which are very strong below the critical temperature. The system will stay in a stable, predominantly negative state for a long period of time, even though the global minimum of the energy landscape is σ^+, because the magnetic field force cannot overcome the barriers between σ^+ and σ^- in the raw Ising energy landscape [27].

3.3. *Attraction-Diffusion and metastability.* Metastability can be observed in any multimodal energy landscape. Let X^* be a local minimum of an energy function U. Even if X^* is a shallow minimum, the temperature can be lowered so that the basin of attraction of X^* is strong enough to trap a local diffusion process. To be more precise, for T less than a critical temperature T_{X^*}, a Markov chain initialized from X^* using a *local* reversible sampling method according to the density

$$p_T(X) = \frac{1}{Z_T} \exp\{-U(X)/T\}$$

will remain trapped in a δ-ball around X^* for a large number of sampling iterations with high probability. The chain becomes trapped in the local mode because the behavior of MCMC is very similar to gradient descent when sampling at low temperature. The acceptance probability for proposals to higher energy regions of the landscape is virtually zero, and any movement away from X^* has high probability of being reversed. To be considered a local sampling method, the probability of displacement in a single step

of the sampler must be virtually 0 above some maximum tolerated step size ε that is small relative to the scale of landscape features. Most standard MCMC methods, such as Random-Walk/Component-Wise Metropolis-Hastings, Gibbs sampling, and Hamiltonian Monte Carlo, are local, or can be tuned to be local.

Now consider two minima X_1^* and X_2^* and suppose $T < \min(T_{X_1^*}, T_{X_2^*})$. Since the diffusion temperature is less than the critical temperature for both minima, an MCMC sample of p_T initiated from either X_1^* or X_2^* should remain in its original basin for a long period of time. Consider the altered density

$$p_{T,\alpha_1,\alpha_2}(X) = \frac{1}{Z_{T,\alpha_1,\alpha_2}} \exp\{-(U(X)/T + \alpha_1||X - X_1^*||_2 + \alpha_2||X - X_2^*||_2)\} \tag{3.4}$$

for magnetization strengths $\alpha_1, \alpha_2 \geq 0$.

Suppose that a sample is initialized from X_2^* according to density $p_{T,\alpha_1,0}$ (i.e., set $\alpha_2 = 0$). If α_1 is sufficiently small, the role of the magnetization term is negligible and the dynamics of the altered distribution are nearly identical to the original distribution. In this case, since $T < T_{X_2^*}$, the sample should remain trapped in the local energy basin of X_2^* and unable to approach X_1^* for a long period of time. On the other hand, it is clear that as $\alpha_1 \to \infty$, X_1^* becomes the sole global minimum of the energy landscape of $p_{T,\alpha_1,0}$ and that an MCMC method initialized from X_2^* would quickly travel to a δ-ball around X_1^* and stay within that ball indefinitely. The same properties hold when the roles of X_1^* and X_2^* are reversed and $\alpha_1 = 0$.

The above observations show that the phase space of p_{T,α_1,α_2} with respect to the non-negative parameters (T, α_1, α_2) in the quarter-planes $(T, \alpha_1, 0)$ and $(T, 0, \alpha_2)$ has properties similar to the phase space of the magnetized Ising energy $U_{T,H}$ with respect to (T, H). The latter model has only two parameters because of the symmetry in the Ising model where $||\sigma - \sigma^+||_1 = 2n - ||\sigma - \sigma^-||_1$, so the magnetization penalties for both σ^+ and σ^- use the same parameter H.

An important difference between the magnetized Ising model and the AD model in a general energy landscape is the asymmetry in the stability of local minima that can occur in the latter case. Detecting asymmetry in the phase space is an essential feature of AD. When the metastable region of one minimum is significantly smaller than the metastable region of the other, this can be evidence that the former minimum belongs to a high-energy region of a large scale funnel, and that the latter minimum is located deeper within the funnel, as in the protein folding model discussed in Section 2.4. See Figure 10 for a practical demonstration of asymmetry in the AD phase space.

The properties of the phase space can be analyzed with local MCMC methods. Such methods are often criticized for their tendency to become trapped in local minima, and for their inability to travel freely throughout the state space. In AD, this "shortcoming" is exploited as a tool for measuring landscape features. When an MCMC sample of U is initiated from a local mode, the correlation over time between the MCMC states and the initial local mode is an *order parameter* that can be used to detect critical phenomena [27]. If temperature is low and an MCMC sample initialized from a local mode is unable to escape, the system is in an *ordered* phase, and the Markov chain remains highly correlated with the local mode indefinitely (i.e., the order parameter remains non-zero).

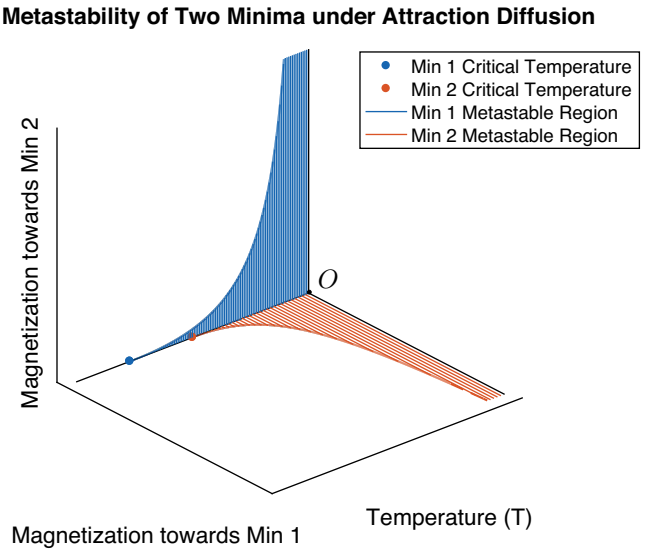


FIG. 9. Metastable regions of the density (3.4) in the parameter space (T, α_1, α_2). The system behavior in the quarter-planes $(T, \alpha_1, 0)$ and $(T, 0, \alpha_2)$ is similar to the upper and lower half of the Ising phase diagram Figure 8, except that the system is not symmetric. The diagram shows that Minimum 1 is more stable, because it has a higher critical temperature and larger metastable region. See Figure 10 for a practical example of this behavior.

When the temperature is high enough to permit escape from the local mode, correlation with the local mode will decay quickly over time (i.e., the order parameter vanishes to 0), representing a *disordered* phase. By examining whether an induced magnetization force disrupts or preserves an ordered phase, it is possible to discover landscape features.

As discussed earlier, a major goal of the present work is to identify macroscopic landscape structures while ignoring noisy local structure. A natural way to accomplish this goal is to shift the focus of the mapping from basins of attraction under gradient descent, the standard practice in ELM applications, to regions of the landscape that are metastable under an MCMC flow, as presented by Bovier [4]. This work divides the landscape into basins where the time-scale of the mixing within basins is exponentially small relative to the time-scale of mixing between basins. Local minima separated only by minor energy barriers belong to the same metastable region. This results in a simple landscape description that directly reflects the dynamics implied by the energy function.

Unfortunately, it is not possible to identify the metastable regions of a landscape simply by initiating MCMC chains from two minima and waiting for the chains to meet, because the "short" time-scales of mixing within basins are far too long for efficient simulation. The magnetization term in AD is meant to accelerate the short mixing time-scales within basins while still respecting the long mixing time-scales between basins. In this way, we can computationally identify the metastable regions described in [4],

because the metastable regions of the magnetized landscape should be very similar to the metastable regions of the original landscape as long as α is not too strong.

In the worst case scenario, for any temperature T, all local minima collapse into a single mode above a threshold α_T, while an essentially infinite number of minima can be found when the magnetization is below α_T. However, if the energy landscape has a manageable number of macroscopic basins, there should be a critical range of (T, α) that will allow movement across the small noisy barriers within metastable basins while restricting movement across the large barriers between basins.

3.4. *Attraction-Diffusion in an image landscape.* We demonstrate the principles of AD using an energy function defined over 16×16 grayscale images of the digits 0, 1, 2, and 3. Each pixel is discretized to 8 values from 0 to 255 and a Gibbs sampler is used for MCMC. In this experiment, we perform AD directly in the 256-dimensional image space. Although the images are small, the number of dimensions is quite large for an ELM application, which typically deal with landscapes of at most 100 dimensions. The energy function and minima are taken from the first experiment in Section 5.3. We trained the network using the DeepFRAME training method in Algorithm 1 with 500 examples of the MNIST digits 0, 1, 2, and 3 each. The energy network structure is given in the second row of Table 1 in the appendix. Minimum A is Minimum 5, Minimum B is Minimum 4, and Minimum C belongs to the group represented by Minimum A in the DG of Figure 19.

The metastable regions of each minima pairing in the parameter space (T, α) can be mapped using AD, and the results are similar to the phase space of the magnetized Ising function, as described in Section 3.2. We used an improvement limit $M = 20$ (one Gibbs sweep is a single iteration) and distance resolution $\delta = 150$ (each pixel has a value from 0 to 255, so this resolution is quite strict). For a range of temperatures spaced evenly on log scale, we estimated the metastable threshold of α by searching for the point where diffusion just failed to reach the target. We started at a high value of α, and attempted 20 AD trials for each pairing. If any of these trials were successful, we decreased the value of α by 3% and ran another 20 trials, and repeated until none of the trials were successful. The minimum energy barrier found during the search was recorded. The minima played both roles in each pairing, so there were 6 tests in total. The plots, shown in Figure 10, validate the AD principles discussed in Section 3.3 and are evidence that the autocorrelation of an MCMC sample can be used as a reliable metric for metastable phenomena in an energy landscape.

Figure 10 also gives an idea of how AD can be used to group minima. The plots show that Minimum C collapses to Minimum A in a region of the parameter space where the other minima are highly stable. Moreover, the barrier found along the AD path between Minimum A and Minimum C is almost 0, despite the fact that the minima are distant in Euclidean space and are separated by an energy barrier along the 1D interpolation path. This is evidence that Minimum C is located along the side of a "funnel" of the energy basin represented by Minimum A, much like an intermediate state in protein folding.

AD can also be used as a method for estimating the MEP between minima. When finding MEP estimates, it is best to run AD chains below critical temperature using a magnetization α that is just above the metastable boundary. Running AD chains at or

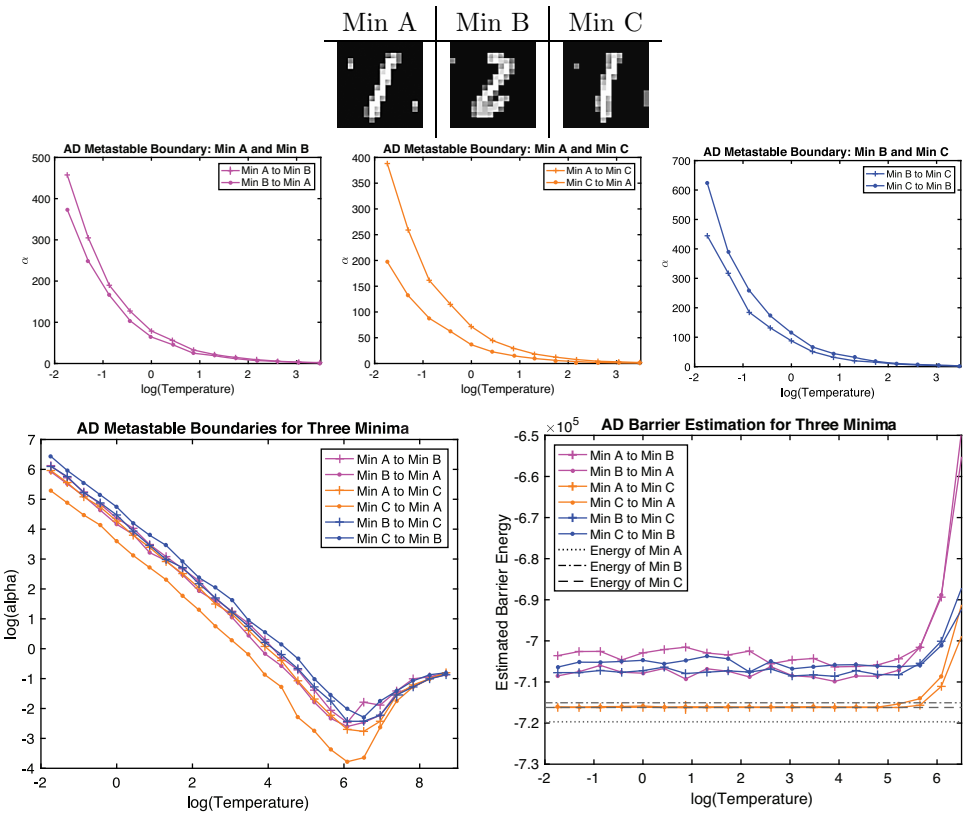


FIG. 10. Top: The three minima tested. Middle: Metastable regions for the minima pairs AB, AC, and BC, respectively. These plots are a superimposition of the two planes from Figure 9. Bottom Left: Comparison of metastable boundaries. Min C merges with Min A at a low α, while the other minima merge at around the same energy level. The relation is approximately linear, and the upward turn reveals the critical temperature. Bottom Right: Barrier estimates across T.

above critical temperature yields poor results because the chains will not be restricted to the lowest-energy regions of the landscape. When α is too strong, the interpolations will be very close to the 1D linear interpolation, because the chain will ignore landscape features and simply travel straight to the target. When α is too low, the chain will never reach the target and no barrier estimate can be obtained. In a small critical region above the metastable boundary, the magnetization force and energy features have equal magnitude and jointly encourage the chain to travel to the target while respecting landscape structure.

Figure 11 shows interpolations performed in a 16×16 image space using the energy network from the first experiment in Section 5.3. The red curve gives the barrier along the 1D linear path between minima in the image space, while the blue curve shows the energy of a successful AD path between the minima. The barriers estimated by AD are

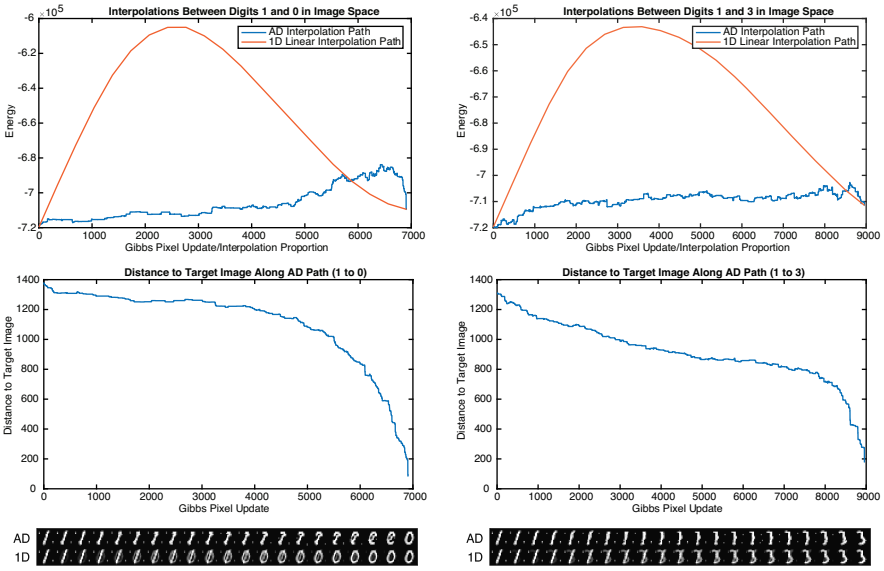


FIG. 11. Top: 1D Interpolation Barrier vs. AD Barrier for diffusion from digit 1 to digit 0, and from digit 1 to digit 3. The AD barriers are much lower, and the AD paths are quite flat. Middle: Distance from target minima vs. Gibbs Sweep. Bottom: Visualization of interpolations. The AD paths are able to move along the image manifold using only an energy network.

drastic reductions of the 1D estimates. Visualizing the images in the AD path shows that the chains diffuse along the image manifold to find non-linear interpolations using only an energy function. AD can also be used to refine pathways in a latent space of a generator network, as we show in Figure 25.

4. Mapping the energy landscape using Attraction-Diffusion.

4.1. *Three essential steps of ELM.* ELM methods have three basic exploration steps:

(1) Get a state X as the starting point for a minima search.
(2) Find a local minimum Y starting from X.
(3) Determine if Y is grouped with a previously found minima basin or if Y starts a new minima basin.

These steps are repeated until no new local minima are found for a certain number of iterations. After the local minima are identified, the barriers between the minima are estimated.

Step 2 can be accomplished with standard gradient descent methods, and the GWL Algorithm provides a principled way to propose X in Step 1. Previous ELM methods lack a reliable way to tackle Step 3. Traditionally, ELM studies have attempted to enumerate *all* basins of attraction of the energy landscape (or the N lowest-energy minima), no matter how shallow [2, 9, 44, 51]. Minima are only grouped together if they are identical

in discrete spaces, or if they are extremely close in continuous spaces. This approach is doomed to failure in all but the simplest cases, because the number of distinct local minima grows exponentially with landscape complexity and/or dimension. On the other hand, for some families of energy functions, the *macroscopic* structure might remain unchanged as landscape complexity/dimension increases. For example, the Ising energy landscape will always have two global basins, regardless of neighborhood structure or number of nodes.

Instead of dividing up the state space according to basins of attraction under gradient flow, we follow the approach of Bovier [4] and divide the state space according to disjoint regions which are metastable under the flow induced by a reversible MCMC process. This results in a much simpler description of the landscape, because the metastable regions will merge basins of attraction which are only separated by small barriers. If the magnetization α used in AD is weak, the metastable regions of the altered landscape should roughly correspond to the metastable regions of the original landscape, and the success or failure of an AD trial can be used as an indicator of membership in a given metastable region. Mapping regions that are metastable under an MCMC process rather than basins of attraction under gradient flow is essential for the success of ELM in complex landscapes.

4.2. *Attraction-Diffusion ELM algorithm.* We now present an Attraction-Diffusion Energy Landscape Mapping (ADELM) Algorithm. Steps 1 and 2 do not involve AD and the implementation details are left open-ended.

The MCMC sampler S should be local in the sense that displacement after a single step is small relative to landscape features with high probability. MCMC methods with step size parameter ε such as Metropolis-Hastings with a Gaussian proposal or HMC/Langevin Dynamics are local samplers, since ε can be tuned to limit displacement. Gibbs sampling is also local, because only a single dimension is changed in each update. The requirement that S is local is needed to ensure that a Markov chain updated using S cannot escape from local modes at low temperatures. Usually, this is considered an undesirable feature of MCMC methods, but in AD it is essential that the Markov samples remain trapped in the absence of magnetization. Upsetting this baseline behavior by introducing a magnetic field enables the discovery of landscape features.

In the ADELM Algorithm, the global minima Z_j of each basin are used as the targets for AD trials. One reason for this choice is the intuition that, for the same strength α, an AD chain should be more likely to successfully travel from a higher-energy minimum to a lower-energy minimum than vice versa. While not true in general, in practice the intuition holds in most cases, especially for very deep minima. A more nuanced implementation could consider multiple candidates from the same basin as targets for diffusion instead of just the global minimum.

Correct tuning of T and α is essential for good results. The temperature T must be set low enough so that movement is restricted to the current mode, but not so low that the chain becomes totally frozen. In our experiments, we first tune the temperature independently of α by initializing unmagnetized chains from a local minimum and observing at the change in energy that occurs over a long trajectory. The change in energy should be small relative to the barriers that exist in the landscape. If the temperature

Algorithm 3: Attraction-Diffusion ELM (ADELM)

input : Target energy U, local MCMC sampler S, temperature $T > 0$, magnetization force $\alpha > 0$, distance resolution $\delta > 0$, improvement limit M, number of iterations N

output: States $\{X_1, \ldots, X_N\}$ with local minima $\{Y_1, \ldots, Y_N\}$, minima group labels $\{l_1, \ldots, l_N\}$, and group global minima $\{Z_1, \ldots, Z_L\}$, where $L = \max\{l_n\}$

for $n = 1 : N$ **do**

1. Get proposal state X_n for minima search. (Random initialization, or a GWL MCMC proposal)
2. Start a local minimum search from X_n and find a local minimum Y_n.
3. **if** $n = 1$, **then**
   Set $Z_1 = Y_1$ and $l_1 = 1$.

else

Determine if Y_n can be grouped with a known group using AD. Let $L_n = \max\{l_1, \ldots, l_{n-1}\}$, and let minimum group membership set $G_n = \emptyset$.

for $j = 1 : L_n$ **do**

a) Set $C = Y_n$, $X^* = Z_j$, $d_1 = ||C - X^*||_2$, $d^* = d_1$, and $m = 0$.

while $(d_1 > \delta)$ & $(m < M)$ **do**

Update C with a single step of sampler S using the density

$$P(X) = \frac{1}{Z} \exp\{-(U(X)/T + \alpha||X - X^*||_2)\}$$

and find the new distance to the target minimum: $d_1 \leftarrow ||C - X^*||_2$.
If $d_1 \geq d^*$, **then** $m \leftarrow m + 1$, **else** $m \leftarrow 0$ and $d^* \leftarrow d_1$.

b) Set $C = Z_j$, $X^* = Y_n$, $d_2 = ||C - X^*||_2$, $d^* = d_1$, and $m = 0$, and repeat the loop in Step a).

c) If $d_1 \leq \delta$ or $d_2 \leq \delta$, then add j to the set G_n, and let B_j be the barrier along the successful path. If both paths are successful, let B_j be the smaller of the two barriers.

if G_n *is empty,* **then**
Y_n starts a new minima group. Set $l_n = \max\{l_1, \ldots, l_{n-1}\} + 1$, and $Z_{l_n} = Y_n$.

else
Y_n belongs to a previous minima group. Set $l_n = \text{argmin}_j B_j$.
if $U(Y_n) < U(Z_{l_n})$, **then**
Update the group global minimum: $Z_{l_n} \leftarrow Y_n$.

is too high, MCMC samples can easily cross between metastable regions even without magnetization and the mapping fails to recover meaningful structure. See Figure 13 for an example of tuning AD temperature.

The magnetization strength α must be strong enough to overcome the noisy shallow barriers in the landscape while respecting the large-scale barriers. Once the temperature T has been tuned and fixed so that chains can diffuse in a limited metastable region, one can run trial mappings across the spectrum of α to locate the critical range where α yields meaningful mapping results. In the limiting case $\alpha \to 0$, each distinct minimum

defines its own metastable region, while in the limiting case $\alpha \to \infty$, all minima merge in a single superbasin. By plotting the number of minima that are discovered in a small number of trial steps as a function of α, it is possible to quickly identify the critical range where magnetization and energy features compete on approximately equal footing. See Figure 13 for an example of tuning AD magnetization. Figure 10 shows that the behavior of AD is quite consistent across a range of T below the critical temperature. Choosing α seems to be the most important tuning decision.

Ideally, in each step of the ADELM Algorithm, diffusion to only one basin representative Z_j should be successful. Successful diffusion to a large number of previously found basins is a sign of poor tuning—in particular, either the value of T or α (or both) is too high, causing leakage between basins. On the other hand, some leakage between minima is usually inevitable, because there are often plateau regions that sit between stronger global basins. This is not too much of a problem as long as the basin representatives remain separated. The global basin representatives $\{Z_j\}$ should be checked periodically to make sure they remain well-separated at the current parameter setting. If an AD chain successfully travels between two of the $\{Z_j\}$, these minima should be consolidated into a single group. This is especially important in the early stages of mapping, when good basin representatives have not yet been found. A single basin can split into multiple groups if the early representatives are not effective attractor states for the entire basin. When consolidating minima, the lower-energy minimum is kept as the group representative.

The ADELM Algorithm has two computational bottlenecks: the local minima search in Step 2, and the AD grouping in Step 3. The computational cost of Step 2 is unavoidable for any ELM method, and the MCMC sampling in Step 3 is not unreasonable as long as it has a comparable running time. In our experiments, we find that the running time for local minimum search and a single AD trial are about the same. Step 3 of the ADELM Algorithm involves AD trials between a new minimum and several known candidates, and the efficiency of ADELM can be greatly increased by running the AD trials in parallel.

4.3. *Barrier estimation and landscape visualization.* AD can be used to estimate the energy barriers and the MEP between local minima after exploration is over. This is done by fixing the temperature T and tuning α to find a threshold where successful travel between minima is just barely possible. The AD barrier estimates are lowest when α is just above the metastable border in the AD phase space, and will increase as α increases. In the limit $\alpha \to \infty$, the AD barriers are identical to the 1D linear barriers, because the MCMC samples will simply move in a straight line towards the target. Estimated barrier height appears consistent for a range of T below critical temperature, as in Figure 10. In our mappings, we are primarily interested in the energy barriers between the global basin representatives, which are the most significant features of the macroscopic landscape.

Disconnectivity Graphs, or DG's (see Section 2.7 and Figure 5), have been used in many previous ELM studies as a method for visualizing the energy landscape. Construction of a DG is straightforward once the minima have been identified by ADELM and the barriers have been estimated by running AD trials between the basin representatives. In our ELM visualizations, we introduce two new elements to the standard DG format. First, we draw circles around the minima nodes of the DG whose size is proportional to

the number of local minima sorted into the corresponding global basin. Second, when mapping image landscapes, we display the global basin representatives in a row at the bottom of the DG, and above the basin representatives, we display randomly selected examples of minima images sorted into each basin, sorted from top to bottom in order of decreasing energy. See Figure 17 for an example.

5. Experiments. We present several experiments that apply the ADELM Algorithm to map the energy landscape of non-convex energy function. The first experiment maps the landscape of an SK spin-glass and compares the ADELM DG to the GWL DG. The remaining experiments focus on the mapping image potentials over both the image and latent space. The AD parameters and network structures used in each experiment can be found in the appendix.

5.1. *Mapping an SK spin-glass.* In our first ADELM experiment, we map the structure of a sample from the 100-state SK spin-glass model. The N-state SK spin-glass is a generalization of the standard N-state Ising model where the coefficients for couplings unspecified. The energy function for the N-state SK spin-glass is

$$U(\sigma) = -\frac{1}{TN} \sum_{1 \le i < k \le N} J_{ik}\, \sigma_i \sigma_k, \tag{5.1}$$

where $\sigma_i = \pm 1$, $T > 0$ is the temperature, and J_{ik} are couplings. In the standard Ising model, the coupling coefficients are either 1 (i.e., the nodes are adjacent) or 0 (i.e., the nodes are not adjacent). The energy landscape of an SK spin-glass contains multiple well-separated global basins that have noisy local structure. Like the Ising model, the landscape is exactly symmetric, since $U(\sigma) = U(-\sigma)$.

Computationally mapping the local minima structure of an SK spin-glass is a challenging task, because an exhaustive search of the state space is infeasible for $N > 30$, and the landscape structure is highly non-convex. Zhou [51] has shown that the GWL Algorithm can accurately identify the lowest-energy minima and barriers for as many as $N = 100$ states. Mapping a 100-dimensional SK spin-glass is a good setting for validating our ADELM Algorithm because the results of our mapping can be compared with the results of a GWL mapping, which are very close to the ground truth. The symmetry of SK spin-glass landscapes is also useful for evaluating our method, because we can compare the mappings of the mirror basins.

We replicated the GWL mappings in [51], and the result is shown in Figure 12. The couplings J_{ik} are independent Gaussians with mean 0 and variance $1/N$, as in the original experiment. We ran our mapping for 5×10^8 iterations using the same GWL parameters described in the original paper, and searched for the 500 lowest minima in the landscape. The number of local minima in an SK spin-glass is far more than 500 even with only $N = 100$ states, but previous mappings show that the 500 lowest-energy local minima capture the main landscape features. In more complex landscapes or larger spin-glasses, even the lowest-energy regions can contain an astronomical number of local minima, making the GWL approach problematic.

After running the GWL mapping, the 500 lowest minima identified were exactly symmetric, meaning that for each minima discovered we also identified its mirror state as a

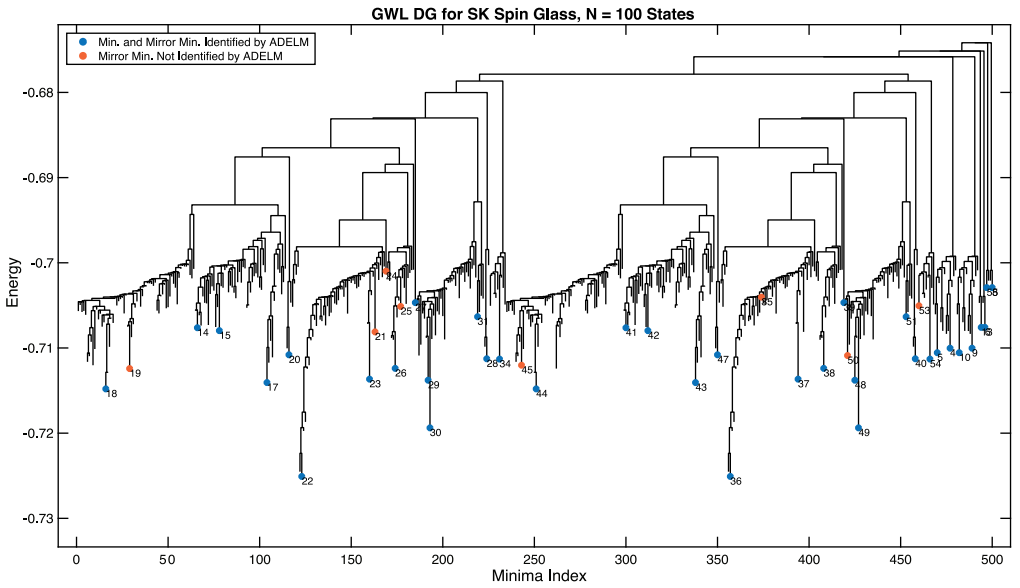


FIG. 12. GWL DG for SK spin-glass. The map records the 500 lowest energy minima in the landscape and the tree is nearly symmetric. The blue and orange dots indicate basins that were identified by our ADELM mapping. The blue dots show minima whose mirror state was also identified during the ADELM run, while the orange dots show minima whose mirror state was not identified. The blue dots cover the most important features of the landscape, which are very stable for the AD parameters ($T = 0.1, \alpha = 1.35$), while the orange dots record substructures within stronger basins that are close to the metastable border.

minima in the mapping. We used two methods to estimate the barriers between minima, and recorded the lower result as the energy barrier. The first method is the one described in [51], which involves identifying transitions between minima basins along the GWL MCMC path and refining these transition states by ridge descent to identify the barrier between the minima.

The second method is a greedy algorithm for interpolation in discrete spaces where we change a starting state to a target state by iteratively choosing, among the spins differing from the target state, the change that causes either the smallest increase or the greatest decrease in energy. Suppose σ and τ are two states, and let $\mathcal{I} = \{i : \sigma_i \neq \tau_i\}$. Let

$$\sigma_j^{(i)} = \begin{cases} \sigma_j & \text{if } j \neq i, \\ \tau_j & \text{if } j = i, \end{cases}$$

for $1 \leq j \leq N$, and $i^* = \operatorname{argmin}_{i \in \mathcal{I}} U(\sigma^{(i)}) - U(\sigma)$. Update the state $\sigma \leftarrow \sigma^{(i^*)}$ and repeat until $\sigma = \tau$. This procedure is not symmetric, so the roles of σ and τ should also be reversed, and the lower barrier of the two paths recorded.

In nearly all cases, the barriers estimated by the second method were significantly lower than the barriers estimated by the first method. Even with the GWL penalty, most MCMC crossings between basins occur well above the minimum energy barrier that

separates the basins. This is corroborated by the observation that the GWL mapping exhibited very poor mixing when we changed the energy spectrum from [-0.8, -0.35], as in the original experiment, to [-0.8, -0.55], which is still well above the maximum barrier between any of the lowest 500 minima. It appears that the global basins of the SK spin-glass model influence the energy landscape in regions that have significantly higher energy than the energy barrier at which the basins merge, and we encounter the same behavior in our other ELM experiments. In this case, it is more appropriate to describe the landscape in terms of metastability, as we are doing in ADELM, rather than barrier height between basins, because the barrier along the MEP is not representative of the energy level at which a diffusion process is affected by a basin of attraction.

We mapped the same energy landscape using ADELM to compare results and to see if ADELM can reliably identify the most important features of the landscape. We used the temperature $T = 0.1$, which is well below the critical temperature $T_c = 1$, and magnetization strength $\alpha = 1.35$ as the AD parameters. Setting T exactly at the critical temperature yielded poor results, because the energy fluctuation of the chains in the absence of magnetization was greater than the depth of the landscape features, and a colder system is needed to restrict diffusion to the lowest energy levels. After tuning and fixing T, we tuned α by running 100 mapping iterations for different α spaced evenly on a log scale and recording the number of minima identified. See Figure 13 for plots showing tuning results. We use the same approach to tune T and α in each of the experiments.

We ran our algorithm for 5,000 iterations, set the AD improvement limit to $M = 100$ Gibbs sweeps of all states, and set our distance resolution $\delta = 0$, which requires that AD chains travel exactly to their target for a successful trial. Our ADELM result is shown in Figure 14, and a side-by-side comparison of the ADELM and GWL mappings is shown in Figure 15. The ADELM mapping identifies the lowest energy minima for all of the major basins of the landscape, as well as substructures within the basins. ADELM is also able to identify a number of basins which are stable but not recorded by the GWL mapping, since these local minima are not among the 500 lowest-energy minima in the landscape. Overall, 44 of the AD basins were also included in the GWL mapping, while 14 stable basins identified by AD were beyond the energy threshold of inclusion in the GWL mapping.

The barriers estimated by the GWL mapping and the ADELM mappings are very similar, although in most cases the GWL barriers are slightly lower than the barriers estimated by AD. This shows that using a large number of minima during barrier estimation can be helpful, because shallow minima can help bridge the gap between stronger basins of attraction. Even though nearly all of the individual barriers identified by GWL are higher than the barriers identified by AD (see Figure 16), the total information of barrier estimates between 500 minima can lead to overall barriers that are lower than the estimates obtained using only 58 minima. On the other hand, it might not be possible to exhaustively identify all of the relevant lowest-energy minima in other landscapes, and it is important to be able to accurately estimate barriers between distant minima without many shallow intermediate minima to connect the basins. Figure 16 shows an AD path between the two global minima of the SK spin-glass. The maximum energy along the

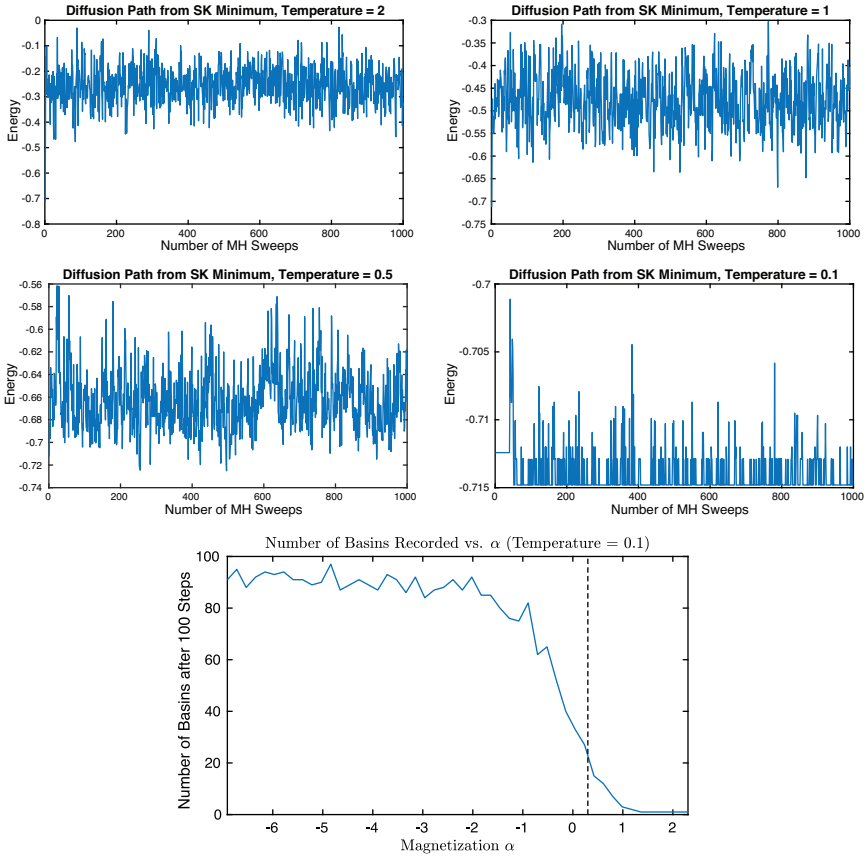


FIG. 13. Top: Tuning the temperature T for AD trials. The system must be cold enough so that MCMC chains do not travel in an energy spectrum above the minimum energy barriers. The critical temperature $T = 1$ is too warm, and we use $T = 0.1$ instead. Bottom: Tuning the magnetization α for AD trials. We run 100 mapping iterations and record the number of distinct basins encountered. As $\alpha \to 0$, we find a new minima for nearly every iteration. As $\alpha \to \infty$, all minima merge into a single basin. In a critical range between the limiting cases, macroscopic behavior can be detected. We use $\alpha = 1.35$, which is shown by the vertical dotted line.

path is only slightly above the barrier identified in GWL and ADELM DG's. This is evidence that AD can provide reliable interpolations between distant locations.

5.2. *Mapping an energy function of image states.* For the rest of our experiments, we use the ADELM Algorithm to map the energy landscape of ConvNet functions which are trained to model real image data. In this section, the target density has the form of the DeepFRAME Model [31, 48]

$$p(I|W) = \frac{1}{Z(W)} \exp\{F(I|W)\}q(I), \tag{5.2}$$

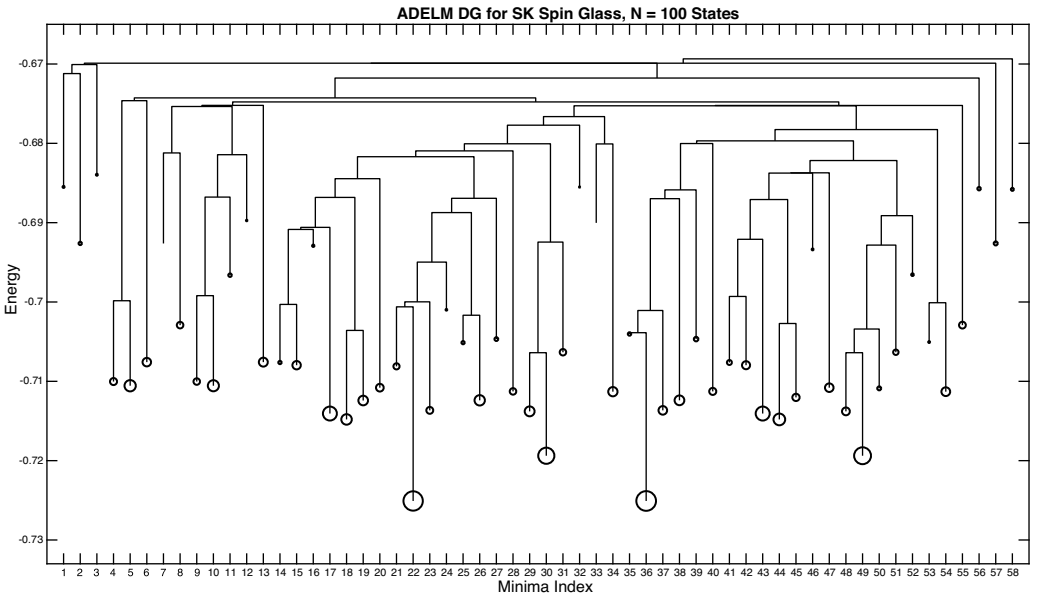


FIG. 14. AD DG for SK spin-glass. The AD diagram is quite symmetric (see Figure 12) and the structure of the DG is very consistent with the DG created from the GWL mapping (see Figure 15). Forty-four of the AD minima are also located by GWL, while 14 of the ADELM minima are not among the 500 lowest energy minima. The GWL mapping, which records only lowest-energy minima, misses significant stable features in higher-energy regions. The size of circles around minima nodes is proportional to the number of minima sorted to each basin, as described in Section 4.3.

where q is the prior distribution $\mathrm{N}(0, \sigma^2 I_D)$, and $F(\cdot|W)$ is a ConvNet function with weights W. The target energy function has the form

$$U(I|W) = -F(I|W) + \frac{1}{2\sigma^2}||I||_2^2. \tag{5.3}$$

All experiments are performed in Matlab using the MatConvNet package [43] for ConvNet implementation.

0-3 *ELM in image space*

In the first ADELM experiment on image models, we map the energy landscape of a DeepFRAME energy directly. The training data are the handwritten digits 0, 1, 2, and 3 from the first half of the MNIST [28] testing set (according to the MNIST documentation, these digits are easier to classify). Each digit has about 500 training examples. The images were resized to 16×16 pixels, and each pixel intensity was discretized to 8 values from 0 to 255. This was done so that a Gibbs sampler could be used as the sampling scheme S in ADELM. We used a Gibbs sampler in this experiment because DeepFRAME landscapes learned with Langevin Dynamics have serious defects in the local mode structure which make mapping impossible. We hope to address this issue in

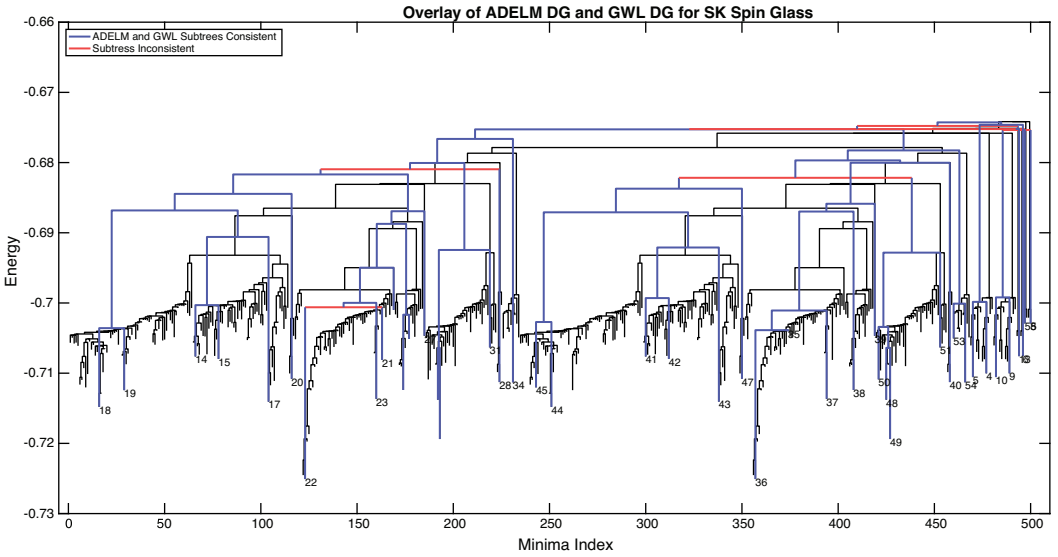


FIG. 15. Overlay of Ising AD and GWL mapping. Blue horizontal lines indicate nodes of the ADELM DG where branch merges are consistent with the GWL DG. Red horizontal lines indicate nodes where the ADELM DG and GWL DG merge branches in a different order. The inconsistencies are minor and mostly occur in higher energy regions. Most inconsistencies only occur for a single merging, and are corrected by the next merge. The ADELM mapping effectively captures the macroscopic features of the GWL mapping.

future work, and eventually we would like to use Langevin Dynamics in the image space as our sampling procedure.

In this experiment, the image space has 256 dimensions. This is larger than the spaces explored in the majority of past ELM experiments, which typically have at most 100 dimensions [2,9,51]. However, use of a Gibbs sampler restricts the size of the image space, since Gibbs sampling scales poorly as dimension increases. We address this problem in later experiments by introducing a generator network, which has a low-dimensional latent space that facilitates movement in the image space of the DeepFRAME energy landscape. Composing a generator network and a DeepFRAME energy provides a way to map the pattern manifold for images of realistic size (see Section 5.4).

The weights W are learned using Algorithm 1. Our training method is the same original method [48] *except* that Gibbs Sampling was used instead of Langevin Dynamics to synthesize images for reasons explained above. The structure of the descriptor network can be found in the appendix. The weights were trained for 300 epochs with a learning rate $\gamma = 0.00007$ and $T = 10$ Gibbs updates of the synthesized images.

We set the improvement limit to $M = 20$ and the distance resolution to 150 (each pixel has intensity between 0 and 255, so about half a pixel). The AD parameters were $T = 30$ and $\alpha = 1.05$. The proposal in Step 1 of the ADELM used random initialization. The mapping was done in two stages: a burn-in stage of 500 iterations, and a testing

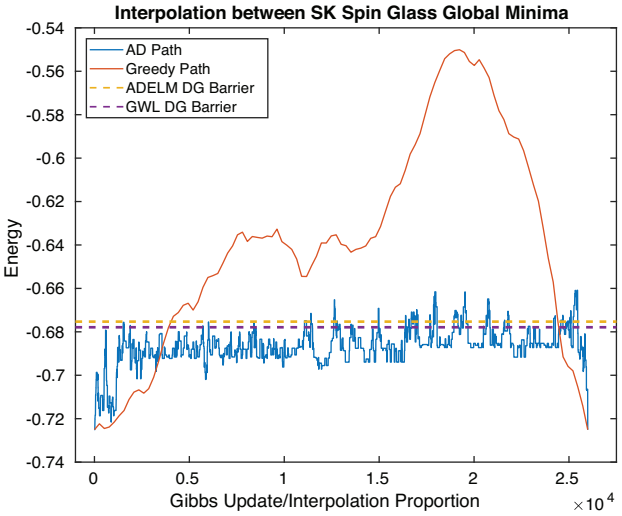


FIG. 16. Interpolation between SK spin-glass global minima. The AD path travels in an energy spectrum very close to the barriers in the GWL and ADELM DG's.

stage of 2000 iterations. After the burn-in stage, the global minima were consolidated by performing AD on all pairs of global minima using the same parameters as during mapping. This is done to weed out extraneous minima that appear early during mapping when good global minima for each basin have not been found. In the testing stage, no new basins were identified, indicating that the mapping procedure has identified the main landscape features.

The ADELM results are shown in Figure 17. The digits 0 and 3 are represented by a single minima, while the digit 1 was split between two basins according to direction of tilt and the digit 2 divided into three groups. We also found two stable basins that do not represent digits. See Section 4.3 for a description of the DG layout.

In this experiment, search for local minima was initiated from white noise images with uniform distribution over each pixel. Figure 18 shows an example path during Gibbs sampling from white noise. The overall pattern of the digit emerges very quickly and the clarity of the digit is slowly refined.

5.3. *Mapping an energy function with generator proposals.* The results in the previous section show that it is possible to map a DeepFRAME energy function by moving through the image space directly. However, many of the local minima identified during mapping are severely distorted digits, or images that do not resemble digits at all. While the DeepFRAME Model builds strong modes that approximate the manifold of the digit data, it also creates many accidental, higher-energy modes that warp the features of the true digit modes. In order to reduce the number of accidental modes discovered in the landscape, one could restrict the proposals in Step 1 of the ADELM algorithm to a region that is close to the true data manifold, instead of using a random image as in the previous section. Since the DeepFRAME energy function has only been trained to

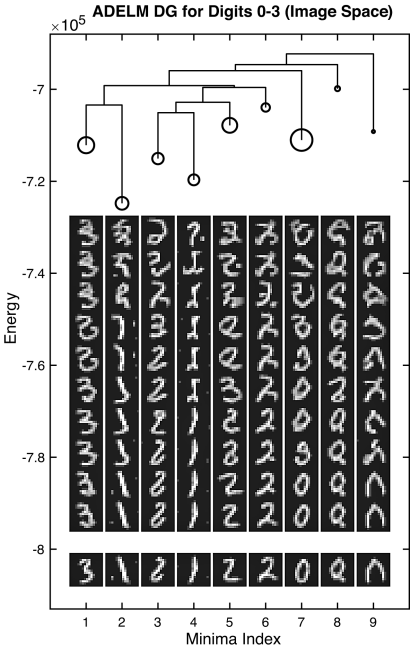


FIG. 17. DG of Digits 0-3 ELM in Image Space. The left-tilted 1 digit and the 3 digit merge at low energy, as do the right-tilted digit 1 and different versions of the digit 2, while the digit 0 remains separate and merges at a higher energy. The images within basins mostly represent the same digit, although many noisy images are identified. The lower-energy images within the basins are well-formed digit images, while the high-energy images are not reliable digit representations. The circles and the organization of the images on the DG are explained in Section 4.3.

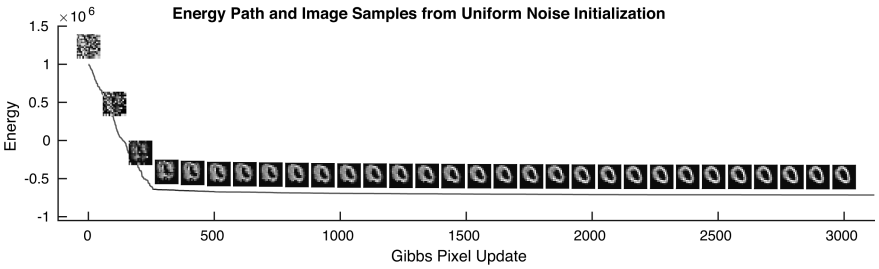


FIG. 18. Local minima search using low temperature Gibbs sampling from a white noise image with a trained DeepFRAME energy.

model a very small subset of the image space (a consequence of any learning algorithm based on Contrastive Divergence), restricting proposals to a region close to the data manifold reveals the structure of the landscape in the regions where the structure is most meaningful.

One way to restrict proposals to a well-formed region of the image landscape is to use a generator ConvNet [14, 20] as the proposal mechanism. As discussed in Section 2.1, an energy function U can be trained jointly with a generator network g using the Co-Op Net Algorithm [49], so it is natural to use g as a proposal mechanism for exploring U. This helps to limit exploration to a region of the image space where U has learned to reliably model the image data. Moreover, mapping the local minima of proposals from the generator network provides a novel way to map the structure of the latent space of the generator networks. Although many authors have made observations about non-linear interpolations in the image space that occur when moving linearly through the latent space, there is no previous work that systematically maps the concepts of a latent space. Following the terminology of [49], we sometimes refer to the DeepFRAME energy U as a *descriptor* network. The descriptor and generator networks are trained using Algorithm 2.

In this section, we only use the generator network to propose new images as a starting point for local minima search. Local minima search and AD are both performed in the 16×16 image space using only an energy network. We extend the role of the generator network further in Section 5.4 by mapping a function of the form $U(Z|W_1, W_2) = U(g(Z|W_2)|W_1)$, where the energy $U(\cdot|W_1)$ is evaluated over the range of $g(\cdot|W_2)$. Since the latent space is much smaller than the image space, this formulation provides a way to efficiently map DeepFRAME energy functions defined over images of realistic size. Interestingly, it appears that the barriers in the landscape of the concatenated energy are more meaningful than the barriers in the raw DeepFRAME landscape, even though the local minima images are very similar (see Figures 21 and 22).

0-3 *ELM in image space with generator proposals*

In our next experiment, we train a Co-Op Network to model the training images of the digits 0, 1, 2, and 3 from the previous section. The generator network structure can be found in the appendix. The learning rate for the generative layer was 0.0003. The descriptor energy was initialized as the energy from Section 5.2 and the learning rate was very low, so the structure of the energy landscape should be similar to that of the previous experiment. Gibbs sampling was used to update the generator images instead of Langevin Dynamics, as discussed in the previous section.

In each iteration of the ADELM Algorithm, we draw a random variable Z from the latent distribution, find the image $g(Z|W_2)$ associated with the latent vector, and use this image as the starting point for local minima search. We used the same ADELM parameters as in Section 5.2. We ran a burn-in sample of 500 iterations, consolidated the minima, and ran a test sample of 2000 iterations to obtain the results shown below.

Figure 19 displays the mapping results. The minima are more consistent with the training data than those found when searching the image space from random initialization. The members of the image basins are coherent, and all basins except for Basin 1 correspond to recognizable digits. The DG structure and basin representatives are very similar to the results in Section 5.2. The DG shows that the basins of the left-tilted 1 and the digit 3 merge at a relatively low energy, and the right-tilted 1 and the skinny 2 merge at a slightly higher energy. Two of the barriers found in the diagram are quite shallow. Nonetheless, these minima are still well-separated under AD at the parameter

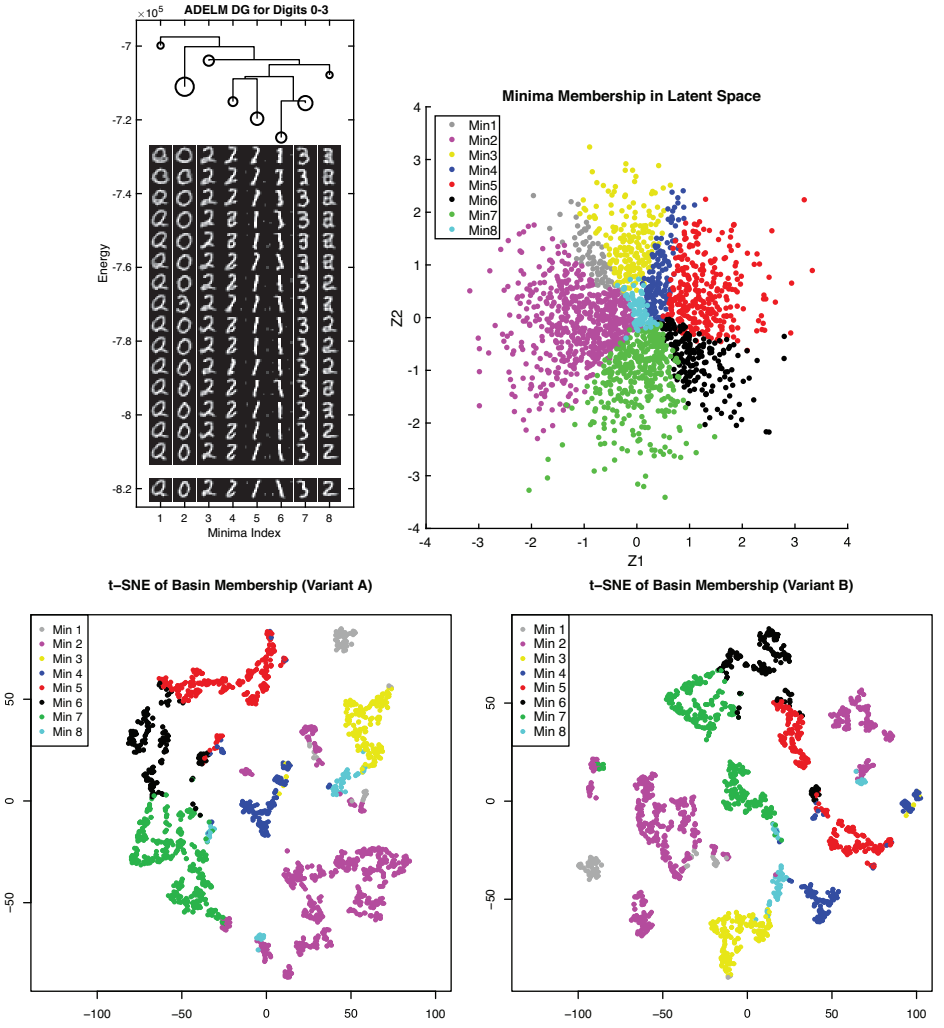


FIG. 19. Top Left: DG for digits 0-3 ELM using generator proposals. Top Right: Latent space $\mathrm{N}(0, I_2)$ colored by basin membership found by ADELM. Bottom: t-SNE visualizations of local minima found in ADELM colored by basin membership.

setting used during mapping. This is evidence that metastability rather than barrier height is best suited for grouping minima, especially if there are stable but flat basins in the landscape. Raw barrier height is not always representative of the dynamics of the system, and global basins influence the landscape well above the energy at which basins merge.

As noted earlier, it is possible to use the ADELM groupings to map the structure of the generator network. Figure 19 shows the latent vectors used to find proposal images, colored according to the ADELM groupings. The ADELM group labels form

well-defined clusters in the latent space, and images representing the same digit are adjacent. Moreover, the arrangement of the latent space reflects the structure of the energy landscape. For example, the group of left-tilted 1's borders the group of the digit 3 in the latent space, and the group of right-tilted 1's borders the group of the skinny 2 digits. We also visualize the minima groupings using t-SNE embeddings. Since t-SNE is a random algorithm, two different results are given. The minima labels match well with the clusters found by t-SNE in both variants.

Spots and Stripes ELM in image space with generator proposals

Next, we map a new descriptor and generator network trained to model small patches from texture images. The textures are shown Figure 20. Five hundred small random patches were taken from each texture image and resized to 16×16 pixels. We used the same network structure and training parameters as in the digits 0-3 ELM experiments, except that the latent space of the generator has 4 dimensions rather than 2. Gibbs sampling was used as the method for updating the synthesized images during training. The AD parameters were $T = 45$ and $\alpha = 1.3$. The other ADELM parameters and procedures were the same as in the previous two experiments.

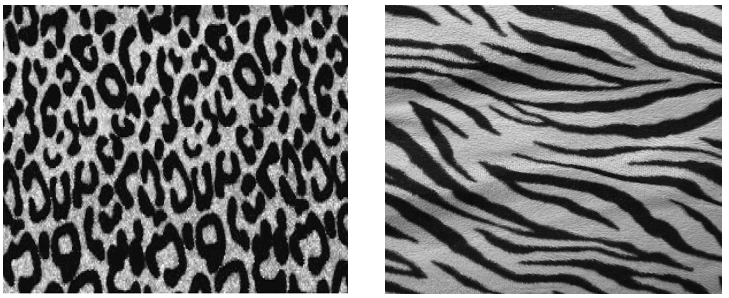


FIG. 20. Spots and Stripes training images. Four hundred random image patches were taken from each image and resized to 16×16 for use as training data.

The results of the Spots and Stripes ELM with generator proposals are shown in Figure 21. Although the appearance of the images within the minima groupings are consistent, the DG has a trivial structure. All minima merge into a single main branch, and the Spots and Stripes do not form separate regions of the energy landscape. This happens because the descriptor landscape has many accidental low-energy regions that are formed as a by-product of CD-style training which obscure the relations between the global basins. Diffusion paths travel through the accidental regions, creating low-energy connections throughout the landscape instead of meaningful barriers. We address this problem in the next section.

5.4. *Mapping energy functions over a latent space.* The ideas of the previous section can be taken a step further by defining an energy function

$$U(Z|W_1, W_2) = U(g(Z|W_2)|W_1) \tag{5.4}$$

over the latent generator variable $Z \in \mathbb{R}^d$, where $U(\cdot|W_1)$ and $g(\cdot|W_2)$ are learned according to the standard Co-Op Net Algorithm. The energy $U(Z|W_1, W_2)$ is very similar

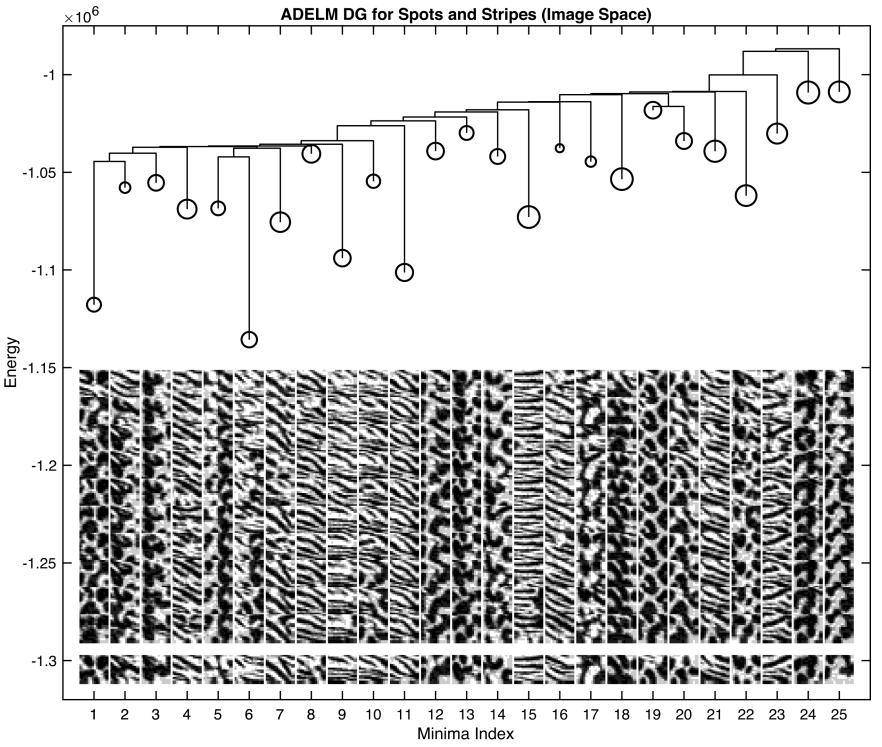


FIG. 21. DG of Spots/Stripes ELM with generator proposals. The tree has a trivial structure where all minima merge along a single branch, and Spots and Stripes cannot be distinguished in the landscape. This happens because the DeepFRAME function creates accidental low-energy regions between modes while it creates the modes. Introducing a generator network helps resolve this problem (see Figure 22).

to the energy used in the DGN-AM model [36], except that we train our generator and descriptor networks jointly to model a dataset of our choosing, while the DGN-AM experiments use a pre-trained GAN for the generator and a pre-trained classification neuron for the descriptor energy. Langevin Dynamics can be used to update the synthesized images during training, because the image space is never sampled directly during AD trials. In previous studies the latent space has a few hundred dimensions at most, and the experiments in Section 5.1 through Section 5.3 show that ADELM can handle such spaces using standard Gibbs sampling or Metropolis-Hastings sampling. The proposal in Step 1 of ADELM can be obtained by sampling from the latent distribution of the generator network. The formulation in (5.4) provides a way to efficiently map DeepFRAME functions defined over images of realistic size using ADELM.

Spots and Stripes ELM in latent space

We use the same Spots and Stripes Co-Op Networks from the previous section and implement ADELM in the 4-dimensional latent space of the generator network to map the energy function (5.4). We use Metropolis-Hastings with Gaussian proposals and a

step size $\varepsilon = 0.025$ as our sampler S, and we set $M = 150$ and $\delta = 0.3$. The AD parameters are $T = 75$ and $\alpha = 300$. We ran 500 burn-in iterations, consolidated the minima, and ran 2000 testing iterations. The proposals in Step 1 of ADELM were drawn from the latent distribution $\mathrm{N}(0, I_4)$. The testing results are shown in Figure 22.

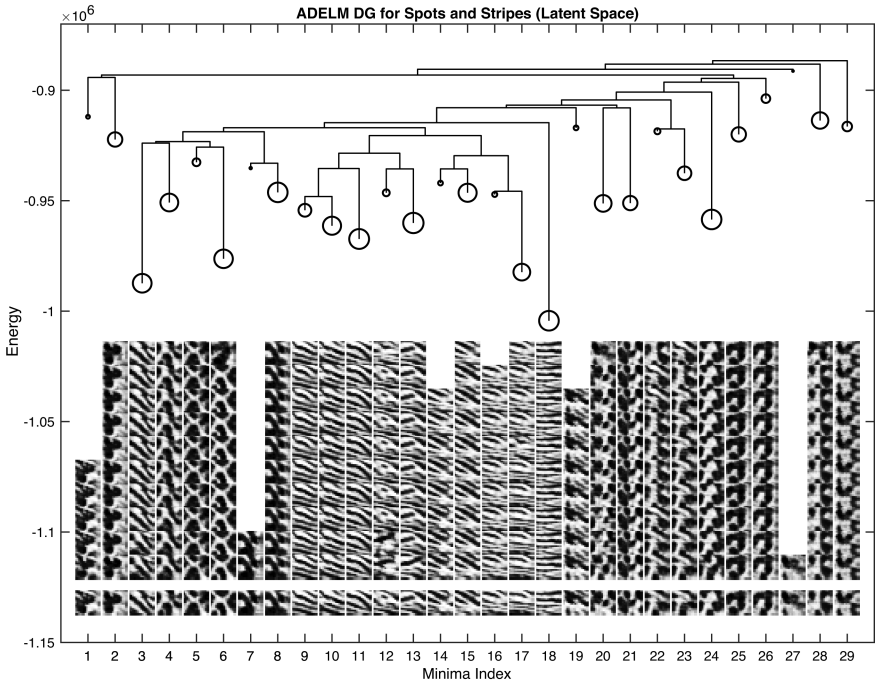


FIG. 22. DG of Spots/Stripes ELM in latent space. The structure is much more complex than the landscape of the same networks mapped using AD directly in the image space, because the generator network has well-defined barriers between its clear images. The barriers generally respect the difference between the Spots and Stripes categories. On the other hand, the Stripes images have lower energy than the Spots images, causing some of the Spots images to merge with the main branch instead of forming their own grouping.

The minima images shown in Figure 22 are very similar to the images in Figure 21, so ADELM recovers about the same global basins in both the image space and latent space. Minima found in the latent space are much more regular than the images from the previous experiment. The latent space DG has a more complex and meaningful structure than the trivial DG from the Spots and Stripes mapping in the image space. Unlike the raw descriptor energy over image space, the joint energy over the latent space contains definite boundaries between the basins of well-formed images.

Figure 22 shows some separation between the Spots and Stripes. Minima 4-8 merge at a low energy and represent Spots (although Minimum 3, an oddball Stripes image,

belongs to the same subtree), while Minima 9-18 are all Stripes. The remaining images are mostly Spots, which merge with these two main subtrees at a higher energy. The algorithm for DG construction is greedy, because branches are merged at the lowest possible energy. This can cause the lower-energy minima (the Stripes) to disrupt the structure among the higher-energy minima (the Spots) in the DG plot. A more nuanced visualization method which groups minima by minimizing the barriers within groups while maximizing the barrier outside of groups in the style of a community-detection algorithm might be able to separate the two categories even more effectively.

Digits 0-9 *ELM in latent space*

Next, we apply ADELM to map the energy (5.4) of Co-Op Networks modeling all of the digits of MNIST. We used the first half of the MNIST testing set as our training data (about 500 examples of each digit). This time, we increase the image size to 64×64 pixels. Since we will only sample in the *latent* space, which has low dimension, we can use realistically-sized images during training.

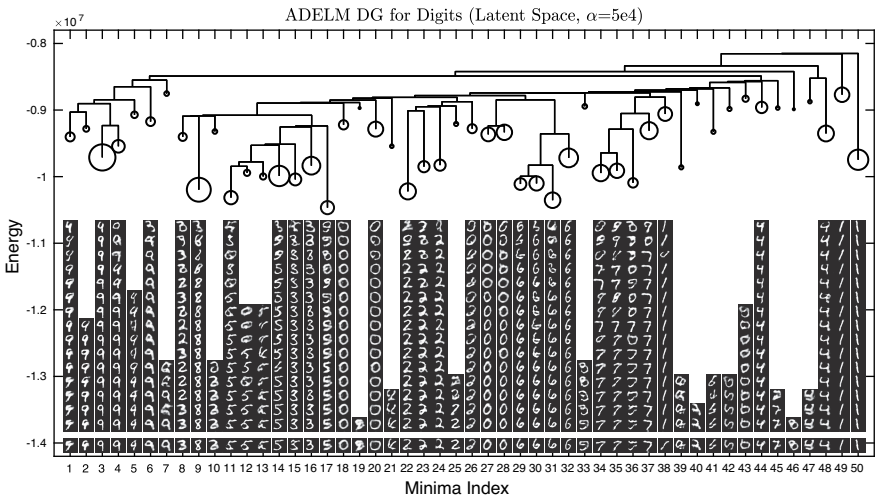


FIG. 23. DG of digits 0-9 ELM in latent space at magnetization α =5e4. The descriptor network is over 64×64 images, but the generator latent space has only 8 dimensions, allowing for efficient mapping. Remarkably, all 10 digits have at least one well-separated branch in the DG. Minima representing the same digit generally merged at low-energy levels.

The descriptor network structure, generator network structure, and AD parameters can be found in the appendix. The other ADELM parameters used were the same as in the Spots/Stripes latent space ELM. For mapping, 500 burn-in iterations and 5000 testing iterations were used, and the results are shown in Figure 23.

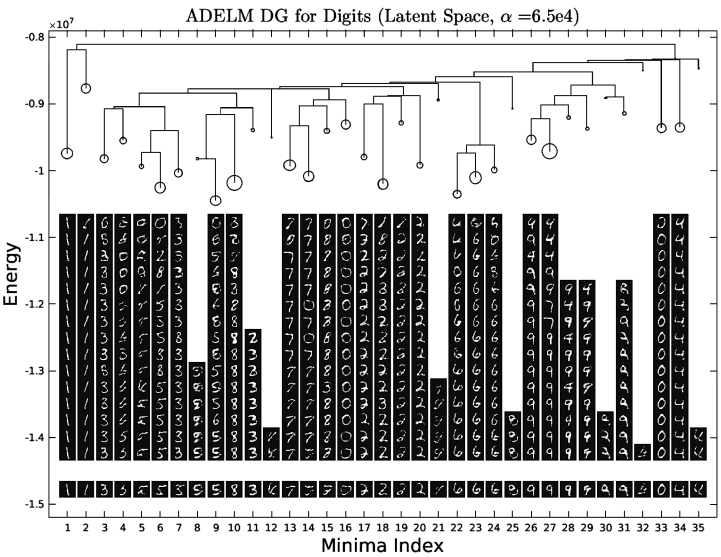


FIG. 24. DG of digits 0-9 ELM in latent space at magnetization α =6.5e4. The same landscape structures appear at different resolution. The basins are not as pure as those found when mapping at a lower magnetization. Some mixing between concepts can be observed, especially in basin 9 and the degenerate image basins 4 and 15.

The DG in Figure 23 has many strong, well-separated energy basins. A close look at the DG shows that all 10 digits are represented by at least a single strong minima basin. The basin members and the global structure of the DG both match closely with human visual intuition. We run a second mapping at a higher magnetization with the same temperature and find many of the same landscape structures. However, the basins of the second mapping are not as pure as the basins of the first mapping and there is some confusion between digit concepts. This indicates that the magnetization strength used in the first mapping is very close to the the maximum magnetization that rigorously preserves concepts that are coherent to a human.

We compare MEP estimates from the DNEB [45] method with MEP estimates from AD. The latent space has only 8 dimensions, so this landscape is a manageable test setting. DNEB uses the 1D linear space between minima as the initial path for further refinement, and Figure 25 shows that the DNEB image paths appear similar to the initial 1D path. On the other hand, the AD paths travel through a different, significantly lower-energy region of the landscape. It is well known that 1D interpolations in the latent space provide more intuitive paths between minima than 1D interpolations in Euclidean space. Figure 25 shows that AD can find interpolations of the latent space that are distinct from the 1D latent interpolation in terms of both energy and appearance. AD and DNEB can be used in conjunction, since AD can provide a rich variety of initialization paths for further refinement by DNEB, which is currently limited to linear 1D initialization.

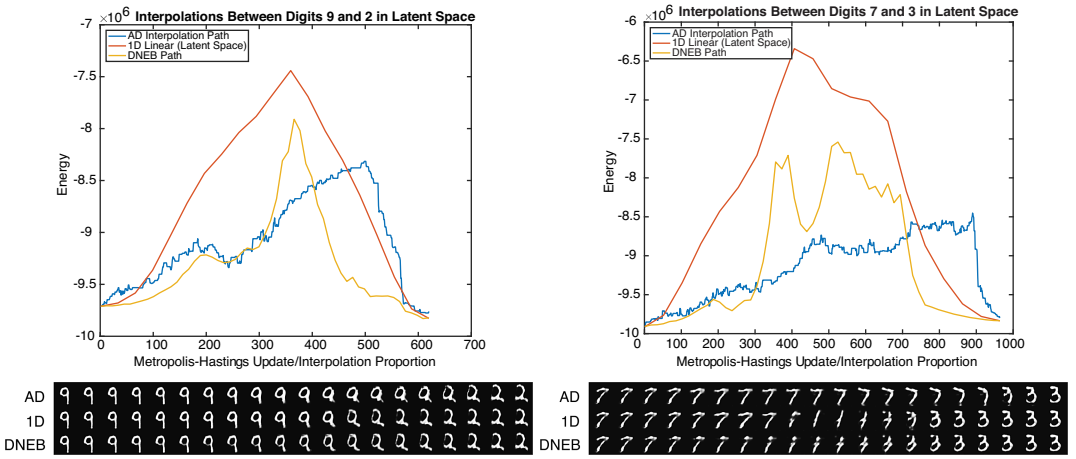


FIG. 25. Comparison of barrier estimation between AD, DNEB [45], and 1D linear interpolation. Top: Barrier estimation with 3 different methods. Both AD paths have lower energy than the 1D linear path and the DNEB path found by refining the 1D path. Bottom: Visualization of interpolations. The DNEB interpolation is almost identical to the 1D interpolation, while AD finds a latent-space interpolation that differs from the 1D linear interpolation in appearance.

Ivy texton ELM in latent space.

We now map a Co-Op Network trained on image patches from an ivy texture. At close range, ivy patches have distinct and recognizable structure, and the goal of the mapping is to identify the main patterns that recur in the ivy textons. Figure 26 shows the entire ivy texture image along with image patches from the texture taken at four different scales. The networks in this experiment are trained to model 1000 image patches from Scale 2.

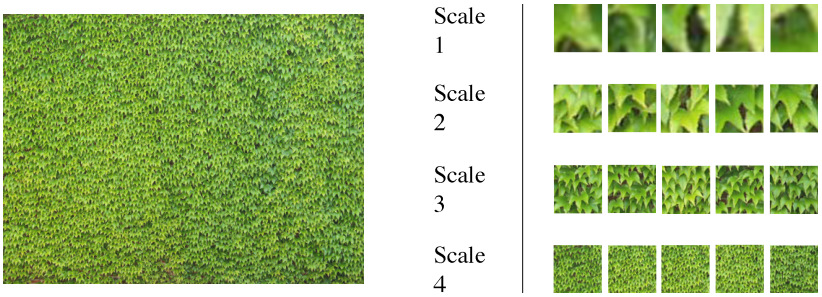


FIG. 26. Ivy texture image and image patches from four scales.

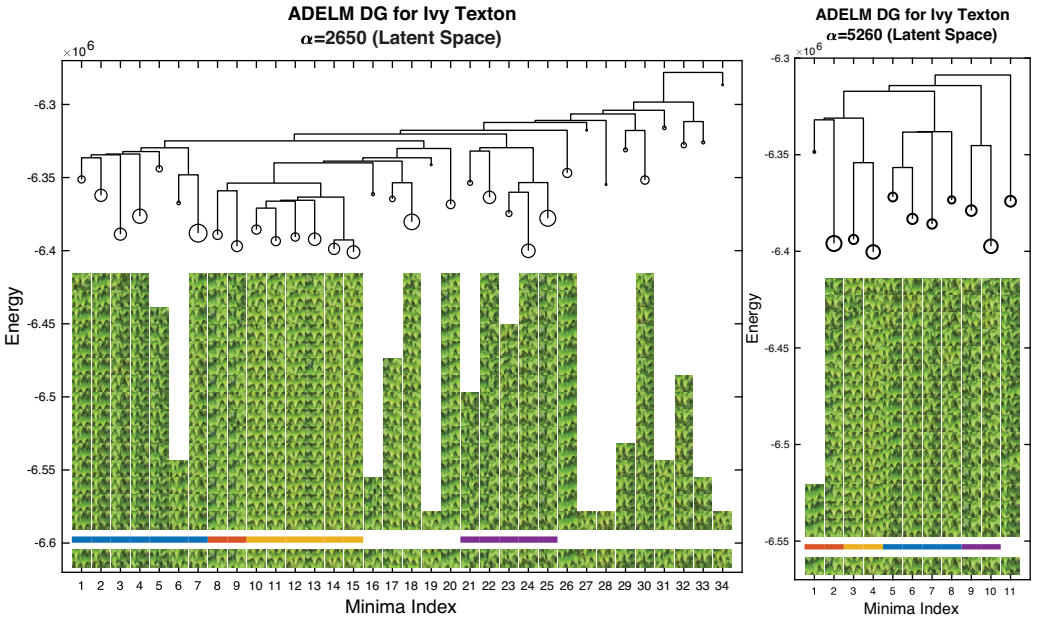


FIG. 27. DG's of ivy textons for two different values of magnetization α. Both mappings show 3 strong global basins and substructures within these basins that are stable at different magnetizations. There is no ground-truth grouping for texton image patches, so it is useful to map image structures at multiple resolutions to identify "concepts" at different degrees of visual similarity. The colors below the basin representatives indicate regions that appear in both mappings.

The DG's for the ivy texton mapping in Figure 27 show that the landscape is dominated by 3 or 4 global basins. The images within basins are very consistent, and the barriers between the basins are representative of visual similarity between the minima images. Unlike the digits mapping, there is no ground-truth for the minima groupings, so it is useful to explore the landscape at different energy resolutions to identify image groupings at different degrees of visual similarity. One major advantage of ADELM is the ability to perform mappings at different energy resolutions simply by changing the magnetization strength α used during the AD trials. Figure 27 presents two mappings of the same landscape at different energy resolutions. The same landscape features appear in both mappings with more or less substructure depending on the magnetization strength.

Multiscale ivy ELM in latent space

We continue our investigation of the ivy texture image from the previous section by mapping a Co-Op Network trained on 1000 image patches from each of the four scales shown in Figure 26. In this experiment, we want to investigate the differences in memory formation between the different scales. In particular, we are interested in identifying a

Ivy Texton $\alpha = 2650$ (Latent Space)

Min. Index	Basin Rep.	Randomly Selected Members (arranged from low to high energy)	Member Count
1			17
3			48
8			29
9			39
10			28
15			53
22			50
25			83

FIG. 28. Minima of ivy texton with magnetization $\alpha = 2650$ in latent space for the DG depicted in Figure 27.

relation between the *metastability* of local minima in the landscape and the *perceptibility* of visual difference among the minima. We expect to find fewer structures at the extreme scales. Image patches from Scale 1 are mostly solid-color images with little variety, which should form a few strong basins in the landscape. Image patches from Scale 4 have no distinct features and cannot be told apart by humans, so we expect these images will form a wide basin without much substructure. For the intermediate scales, we expect to find a richer assortment of stable local minima, because the intermediate scales contain more variation than Scale 1, but the variation still can be distinguished visually, in contrast to the Scale 4 images.

Figure 3 shows the results of our mapping, and Figure 4 gives a closer look at basins from each scale. The structure of the landscape does indeed differ between the image scales. As expected, the memories from Scale 1 form a few strong and large basins. Scale 2 accounts for the majority of the basins in the landscape, since this scale contains the most variety of perceptible image appearances. The Scale 2 basins merge with the Scale 1 basins in the DG visualization, indicating that there are accessible low-energy connections between these regions of the landscape. The images from Scale 3 and Scale 4 each form a separate region of the energy landscape with little substructure. The mapping shows that the perceptibility threshold for ivy texture images (at least in terms of memories learned by the Co-Op Network) lies somewhere between Scale 2 and Scale 3. Above the perceptibility threshold, the network cannot reliably distinguish variation between images, and the landscape forms a single region with no significant substructure. It is difficult for a human to distinguish groups among images from Scale 3, so the perceptibility threshold for the network seems similar to that of humans.

Cat faces ELM in latent space

For our final experiment, we map a Co-Op Network trained on aligned cat face images gathered from the internet. The results of our mapping are shown in Figure 29, and Figure 30 gives a closer look at some of the basins. The DG has a single branch and the energy barriers are quite shallow. The main features of the local minima are the geometry and color of the cat faces, but these can be smoothly deformed during interpolation without encountering improbable images, in contrast to images such as digits, which must enter an improbable geometric configuration along an interpolation path. For this reason, the energy barriers throughout the cat landscape are very low. Nonetheless, the global basins found by ADELM coherently identify major groups of cat faces. AD can effectively identify landscape structure even when the majority of basin members have energy that is higher than the barrier at which the basin merges. This is further evidence that macroscopic basins influence the energy landscape in regions well above the lowest barrier between basins, and that metastability is a more suitable criterion than barrier height for identifying landscape structure.

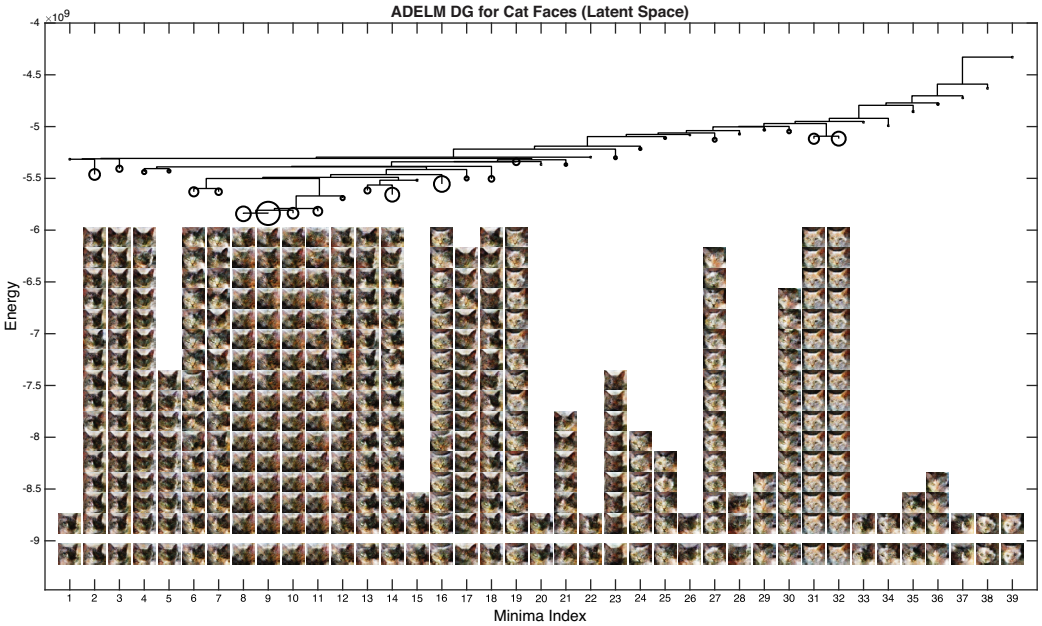


FIG. 29. DG of cat faces in latent space. The landscape has a single global basin, likely because interpolations between cat faces that respect geometry and color constraints are easily found, unlike interpolations between digits, which must pass through a high-energy geometric configuration along the path. Despite the lack of overall landscape structure, AD is able to find meaningful image basins that show a variety of cat faces.

Cat Faces (Latent Space)

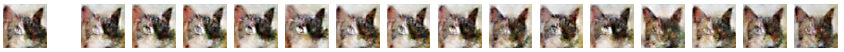
Min. Index	Basin Rep.	Randomly Selected Members (arranged from low- to high-energy)	Member Count
2			157
4			25
6			114
9			729
16			335
18			44
32			263

FIG. 30. Minima of cat faces in latent space for the DG depicted in Figure 29. Despite the shallow barriers found throughout the landscape, there are still metastable regions that capture consistent appearances.

6. Conclusion. This work introduces a new MCMC tool called Attraction-Diffusion, which uses local sampling in an altered landscape to gain information about the relative stability of local minima in the original energy landscape. A unique feature of AD is the exploitation of the high autocorrelation that occurs when MCMC samples are trapped in local modes. In most MCMC research, this phenomenon is considered a major obstacle, but our work uses this aspect of MCMC sampling to measure landscape features. AD learns from both the success and failure of a local Markov sample as the chain is encouraged to escape from local barriers by an induced magnetization. The principles of AD can be traced back to magnetized energy functions from statistical physics, and AD can be interpreted as a way of measuring the metastable regions in the phase space (T, α).

We also introduce a new Energy Landscape Mapping algorithm called ADELM, which uses AD to sort local minima into separate metastable regions. By tuning the AD parameters to permit successful travel across the low-energy barriers *within* metastable basins while respecting the large barriers *between* metastable basins, it is possible to efficiently map the macroscopic structure of complex landscapes with noisy local structure. ADELM convergence is usually quite fast—in the experiments presented, the main energy basins were identified within the first 100 iterations, and the mappings require only a few thousand iterations, whereas previous ELM methods require millions or billions of iterations. AD can also find energy barriers between minima that are lower than the barriers obtained from widely-used MEP estimation methods such as DNEB. The ADELM algorithm can be applied to a wide variety of continuous and discrete energy functions.

Using the ADELM Algorithm, we present a novel ELM application—mapping the local minima structure of ConvNet functions which are trained to model real image data. Our experiments on Gibbs distributions defined by ConvNet functions show that it is possible to computationally identify image memories of a learned density, and that the structure of memories varies according to the images in the training set. The metastable basins identified by ADELM contain coherent groups of images, and the landscape structure of different image patterns reflects aspects of human visual intuition. Our mappings support the conjecture that the metastability of local minima is related to the perceptibility of differences between minima. The memory landscape forms many separate and stable basins when it is able to distinguish variation between low-entropy images, while large basins with little substructure are formed for memories of high-entropy images such as textures.

In future work, we plan to continue mapping the local minima structure of a wide variety of image densities. Although we encountered difficulties when directly mapping energy functions of realistically-sized image spaces, and overcame this by introducing a generator network with a low-dimensional latent space, we hope to eventually perform mapping using only an energy function over the image space. Energy functions trained using a CD-style algorithm develop serious degeneracies in regions of the image space that are distant from the pattern manifold, creating vast accidental low-energy basins that make mapping impossible. We hope to overcome this problem by using an ensemble of energy functions or energy functions at multiple scales to eliminate the accidental low-energy regions found in a single energy function. In the long term, we want to extend our method to identify hierarchical relations between image memories at different scales, and hope to define compositional "dictionaries" that describe how image patches of smaller scales combine to form image patches of larger scales. ADELM shows great potential for future application to many other non-convex energy functions, including statistical loss functions and potential functions of physical systems.

Appendix A. Table of experiment parameters. Table 1 contains image, network, and AD parameters used in each of the mapping experiments. The network structures are presented in the format $layer : (channels\ out, kernel\ size, stride)$. All layers are followed by a ReLU activation function except for the final layer of the generator networks, which use a tanh activation. The descriptor networks use convolutional layers, and the generator networks use convolutional transpose layers. All layers use $padding = \text{floor}([kernel\ size]/2)$ except for fully connected layers and the first layer of generator networks, which use $padding = 0$.

TABLE 1. Experiment Parameters

Experiment	Space Dim.	ELM Params.	Descriptor Network	Generator Network
SK Spin-Glass	100	$\alpha = 1.35$ $T = 0.1$	–	–
Digits 0-3 (Image Space)	Im: 16^2	$\alpha = 1.05$ $T = 30$	1: (50, 5, 2) 2: (100, fully, –) 3: (1, fully, –)	–
Digits 0-3 (Gen. Prop.)	Im: 16^2 Z: 2	$\alpha = 30$ $T = 1.05$	1: (50, 5, 2) 2: (100, fully, –) 3: (1, fully, –)	1: (100, 4, 1) 2: (50, 5, 2) 3: (3, 5, 2)
Spots/Stripes (Gen. Prop.)	Im: 16^2 Z: 4	$\alpha = 45$ $T = 1.3$	1: (50, 5, 2) 2: (100, fully, –) 3: (1, fully, –)	1: (100, 4, 1) 2: (50, 5, 2) 3: (3, 5, 2)
Spots/Stripes (Latent Space)	Im: 16^2 Z: 4	$\alpha = 300$ $T = 75$	1: (50, 5, 2) 2: (100, fully, –) 3: (1, fully, –)	1: (100, 4, 1) 2: (50, 5, 2) 3: (3, 5, 2)
Digits 0-9 (Latent Space)	Im: 64^2 Z: 8	$\alpha_1 = 5e4$ $\alpha_2 = 6.5e4$ $T = 1200$	1: (100, 5, 2) 2: (200, 5, 2) 3: (1, fully, –)	1: (100, 4, 1) 2: (50, 7, 4) 3: (3, 7, 4)
Ivy Texton (Latent Space)	Im: $32^2 \times 3$ Z: 30	$\alpha_1 = 2650$ $\alpha_2 = 5260$ $T = 500$	1: (100, 5, 2) 2: (200, 5, 2) 3: (1, fully, –)	1: (200, 4, 1) 2: (100, 5, 2) 3: (50, 5, 2) 4: (3, 5, 2)
Multiscale Ivy (Latent Space)	Im: $64^2 \times 3$ Z: 30	$\alpha = 3.3e4$ $T = 750$	1: (100, 5, 2) 2: (200, 5, 2) 3: (1, fully, –)	1: (400, 4, 1) 2: (200, 5, 2) 3: (100, 5, 2) 4: (50, 5, 2) 5: (3, 5, 2)
Cat Faces (Latent Space)	Im: $64^2 \times 3$ Z: 30	$\alpha = 1.5e5$ $T = 1500$	1: (100, 5, 2) 2: (200, 5, 2) 3: (1, fully, –)	1: (400, 4, 1) 2: (200, 5, 2) 3: (100, 5, 2) 4: (50, 5, 2) 5: (3, 5, 2)

References

[1] Y. F. Atchadé and J. S. Liu, *The Wang-Landau algorithm in general state spaces: applications and convergence analysis*, Statist. Sinica **20** (2010), no. 1, 209–233. MR2640691

[2] A. J. Ballard, J. D. Stevenson, R. Das, and D. J. Wales, *Energy landscapes for a machine learning application to series data*, Journal of Chemical Physics **144** (2016), 124119.

[3] O. M. Becker and M. Karplus, *The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics*, Journal of Chemical Physics **106** (1997), no. 4.

[4] A. Bovier and F. den Hollander, *Metastability: A potential theoretic approach*, International Congress of Mathematicians **3** (2006), 499–518.

[5] C. J. Cerjam and W. H. Miller, *On finding transition states*, The Journal of chemical physics **75** (1981), no. 2800.

[6] P. Chaudhari, A. Choromanska, S. Soatto, Y. A. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, *Entropy-sgd: Biasing gradient descent into wide valleys*, ICLR (2017).

[7] P. Chaudhari and S. Soatto, *On the energy landscape of deep networks*, arXiv:1511.06485 (2015).

[8] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. A. LeCun, *The loss surfaces of multilayer networks*, AISTATS (2015).

[9] R. Das and D. J. Wales, *Machine learning landscapes and predictions for patient outcomes*, R. Soc. Open Sci. **4** (2017), no. 7, July, 170175, 19, DOI 10.1098/rsos.170175. MR3688315

[10] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, *Visualizing higher-layer features of a deep network*, Technical Report, Univerisite de Montreal (2009).

[11] R. P. Feynman and P. G. Wolynes, *Quantum mechanics and path integrals*, McGraw Hill, New York, 1965.

[12] S. Geman and C.-R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim. **24** (1986), no. 5, 1031–1043, DOI 10.1137/0324060. MR854068

[13] S. German and D. German, *Stochastic relaxation, gibbs distribution, and the bayesian restoration of images.*, IEEE Trans. PAMI **6** (1984), 721–741.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, Advances in Neural Information Processing Systems (2014), 2672–2680.

[15] U. Grenander, *General pattern theory: A mathematical study of regular structures*, Oxford Science Publications, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1993. MR1270904

[16] U. Grenander and M. I. Miller, *Representations of knowledge in complex systems*, with discussion and a reply by the authors, J. Roy. Statist. Soc. Ser. B **56** (1994), no. 4, 549–603. MR1293234

[17] U. Grenander, *Probability models for clutter in natural images*, IEEE Trans. Pattern Analysis and Machine Learning **23** (2001), no. 4.

[18] U. Grenander and M. I. Miller, *Pattern theory: from representation to inference*, Oxford University Press, Oxford, 2007. MR2285439

[19] T. A. Halgren and W. N. Lipscomb, *The synchronous-transit method for determining reaction pathways and locating molecular transition states*, Chemical Physics Letters **49** (1977), no. 2, 225–232.

[20] T. Han, Y. Lu, S.-C. Zhu, and Y. N. Wu, *Alternating back-propagation for generator network*, arXiv:1606.08571 (2016).

[21] G. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation (2002), 1771–1800.

[22] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Nat. Acad. Sci. U.S.A. **79** (1982), no. 8, 2554–2558, DOI 10.1073/pnas.79.8.2554. MR652033

[23] B. Julesz, *Visual pattern discrimination*, IRE Trans. Information Theory **8** (1962), no. 2, 84–92.

[24] B. Julesz, *Textons, the elements of texture perception, and their interactions*, Nature **290** (1981), 91.

[25] A. Krizhevsky, I. Sutskever, and G. Hinton, *Imagenet classification with deep convolutional neural networks*, NIPS (2012), 1097–1105.

[26] A. Kuki and P. G. Wolynes, *Electron tunneling paths in proteins*, Science **236** (1986), no. 1647.

[27] D. P. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics*, 3rd ed., Cambridge University Press, Cambridge, 2009. MR2559932

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE **86** (1998), no. 11, 2278–2324.
[29] F. Liang, *A generalized Wang-Landau algorithm for Monte Carlo computation*, J. Amer. Statist. Assoc. **100** (2005), no. 472, 1311–1327, DOI 10.1198/016214505000000259. MR2236444
[30] P.-L. Loh and M. J. Wainwright, *Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima*, J. Mach. Learn. Res. **16** (2015), 559–616. MR3335800
[31] Y. Lu, S. C. Zhu, and Y. N. Wu, *Learning frame models using cnn filters*, Thirtieth AAAI Conference on Artificial Intelligence (2016).
[32] A. Mahendran and A. Vedaldi, *Visualizing deep convolutional neural networks using natural pre-images*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015), 5188–5196.
[33] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, *On the number of linear regions of deep neural networks*, Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2924–2932.
[34] A. Mordvintsev, C. Olah, and M. Tyka, *Inceptionism: Going deeper into neural networks*, Google Research Blog (2015).
[35] R. M. Neal, *MCMC using Hamiltonian dynamics*, Handbook of Markov chain Monte Carlo, Chapman & Hall/CRC Handb. Mod. Stat. Methods, CRC Press, Boca Raton, FL, 2011, pp. 113–162. MR2858447
[36] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, *Synthesizing the preferred inputs for neuron in neural networks via deep generator networks*, NIPS (2016).
[37] J. N. Onuchic and P. G. Wolynes, *Theory of protein folding*, Current Opinion in Structural Biology **14** (2004), 70–75.
[38] M. Pavlovskaia, K. Tu, and S.-C. Zhu, *Mapping the energy landscape of non-convex learning problems*, arXiv preprint arXiv:1410.0576 (2014).
[39] Z. Si, H. Gong, S.-C. Zhu, and Y. N. Wu, *Learning active basis models by EM-type algorithms*, Statist. Sci. **25** (2010), no. 4, 458–475, DOI 10.1214/09-STS281. MR2807764
[40] J. Simons, P. Joergensen, H. Taylor, and J. Ozment, *Walking on potential energy surfaces*, Journal of Physical Chemistry **89** (1985), no. 684.
[41] T. Tieleman, *Training restricted boltzmann machines using approximations to the likelihood gradient*, ICML (2008), 1064–1071.
[42] L. Van der Maaten and G. Hinton, *Visualizing data using t-sne*, Journal of Machine Learning Research **9** (2008), no. 85, 2579–2605.
[43] A. Vedaldi, K. Lenc, and G. Ankush, *Matconvnet – convolutional neural networks for matlab*, Proceeding of the ACM Int. Conf. on Multimedia (2015).
[44] D. J. Wales, *The energy landscape as a unifying theme in molecular science*, Phil. Trans. R. Soc. A **363** (2005), 357–377.
[45] D. J. Wales and S. A. Trygubenko, *A doubly nudged elastic band method for finding transition states*, Journal of Chemical Physics **120** (2004), 2082–2094.
[46] F. Wang and D. P. Landau, *Efficient multiple-range random walk algorithm to calculate the density of states*, Physical review letters **86** (2001), 2050–2053.
[47] Y. N. Wu, C.-E. Guo, and S.-C. Zhu, *From information scaling of natural images to regimes of statistical models*, Quart. Appl. Math. **66** (2008), no. 1, 81–122, DOI 10.1090/S0033-569X-07-01063-2. MR2396653
[48] J. Xie, W. Hu, S. C. Zhu, and Y. N. Wu, *A theory of generative convnet*, International Conference on Machine Learning (2016).
[49] J. Xie, Y. Lu, and Y. N. Wu, *Cooperative learning of energy-based model and latent variable model via mcmc teaching*, AAAI (2018).
[50] Y. Zeng, X. Penghao, and G. Henkelman, *Unification of algorithms of minimum mode optimization*, Journal of Chemical Physics **140** (2014), 044115.
[51] Q. Zhou, *Random walk over basins of attraction to construct ising energy landscapes*, Physical Review Letters **106** (2011), 180602.
[52] S.-C. Zhu, X. Liu, and Y. N. Wu, *Exploring texture ensembles by efficient markov chain monte-carlo*, PAMI **22** (2000), 245–261.
[53] S.-C. Zhu, Y. N. Wu, and D. Mumford, *Filters, random fields and maximum entropy (frame): Toward a unified theory for texture modeling*, International Journal of Computer Vision **27** (1998), no. 2, 107–126.