Decoding Stacked Denoising Autoencoders

Sho Sonoda* Noboru Murata
Faculty of Science and Engineering
Waseda University

May 9, 2016

Abstract

Data representation in a stacked denoising autoencoder is investigated. *Decoding* is a simple technique for translating a stacked denoising autoencoder into a composition of denoising autoencoders in the ground space. In the infinitesimal limit, a composition of denoising autoencoders is reduced to a *continuous denoising autoencoder*, which is rich in analytic properties and geometric interpretation. For example, the continuous denoising autoencoder solves the *backward heat equation* and transports each data point so as to decrease entropy of the data distribution. Together with ridgelet analysis, an *integral representation of a stacked denoising autoencoder* is derived.

1 Introduction

The *denoising autoencoder (DAE)* is a role model for representation learning, the objective of which is to capture a good representation of the data. Vincent et al. [2008] introduced it as a heuristic modification of traditional autoencoders for enhancing robustness. In the setting of traditional autoencoders, we train a neural network as an identity map $X \mapsto X$ and extract the hidden layer to obtain the so-called "code." On the other hand, the DAE is trained as a denoising map $\widetilde{X} \mapsto X$ of deliberately corrupted inputs $\widetilde{X}$. The *corrupt and denoise* principle is simple, but truly is compatible with stacking, and thus, inspired many new autoencoders. See Section 1.1 for details.

We are interested in *what deeper layers represent* and *why we should deepen layers*. In contrast to the rapid development in its application, the stacked autoencoder remains unexplained analytically, because generative models, or probabilistic alternatives, are currently attracting more attention. In addition, deterministic approaches, such as kernel analysis and signal processing, tend to focus on convolution networks from a group invariance aspect. We address these questions from deterministic viewpoints: transportation theory and ridgelet analysis.

*s.sonoda0110@toki.waseda.jp

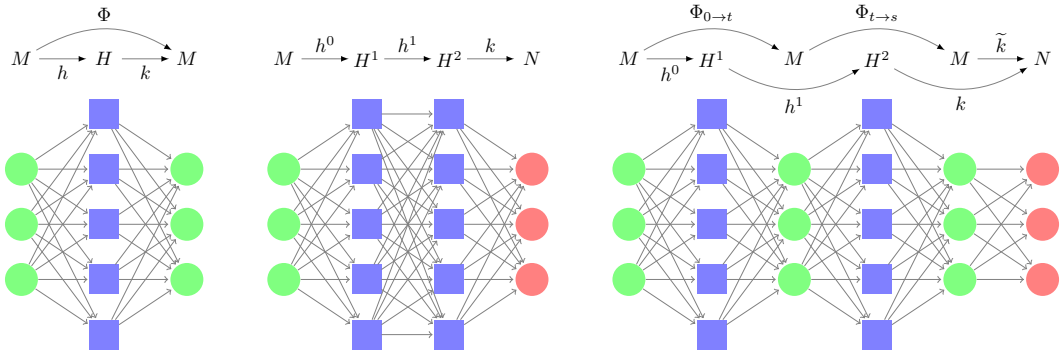


Figure 1: Denoising autoencoder (**left**), stacked denoising autoencoder with linear output (**center**), and a composition of two denoising autoencoders (**right**). Decoding translates a stacked denoising autoencoder into a composition of denoising autoencoders.

Alain and Bengio [2014] derived an explicit map that a shallow DAE learns as

$$x \mapsto \frac{\mathbb{E}_\varepsilon[p(x-\varepsilon)(x-\varepsilon)]}{\mathbb{E}_\varepsilon[p(x-\varepsilon)]}, \tag{1}$$

and showed that it converges to the *score* $\nabla \log p$ of the data distribution p as the variance of ε tends to zero. Then, they recast it as manifold learning and score matching. We reinterpret (1) as a *transportation map* of x, the variance as *time*, and the infinitesimal limit as the initial velocity field.

Ridgelet analysis is an integral representation theory of neural networks [Sonoda and Murata, 2015, 2014, Candès, 1998, Murata, 1996]. It has a concrete geometric interpretation as wavelet analysis in the Radon domain. We can clearly state that the first hidden layer of a stacked DAE is simply a discretization of the ridgelet transform of (1). On the other hand, the character of deeper layers is still unclear, because the ridgelet transform on stacked layers means the composition of ridgelet transforms $\mathscr{R} \circ \mathscr{R}$, which lacks geometric interpretation. One of the challenges here is to develop the integral representation of deep neural networks.

We make two important observations. First, through *decoding*, a stacked DAE is equivalent to a composition of DAEs. By definition, they differ from each other, because "stacked" means a concatenation of autoencoders with each output layer removed, while "composition" means a concatenation of autoencoders with each output layer remaining. Nevertheless, decoding relates the stacked DAE and the composition of DAEs. Then, ridgelet transform is reasonable, because it can be performed layer-wise, which leads to the integral representation of a deep neural network.

Second, an infinite composition results in a *continuous DAE*, which is rich in analytic properties and geometric interpretation, because it solves the *backward heat equation*. This means that what deep layers do is to transport mass so as to decrease entropy. Together with ridgelet analysis, we can conclude that what a deep layer represents is a discretization of the ridgelet transform of the transportation map.

1.1 Related Work

Vincent et al. [2008] introduced the DAE as a modification of traditional autoencoders. While the traditional autoencoder is trained as an identity map $X \mapsto X$, the DAE is trained as a denoising map $\widetilde{X} \mapsto X$ for artificially corrupted inputs $\widetilde{X}$, in order to enhance robustness.

Theoretical justifications and extensions follow from at least five aspects: manifold learning [Rifai et al., 2011, Alain and Bengio, 2014], generative modeling [Vincent et al., 2010, Bengio et al., 2013, 2014], infomax principle [Vincent et al., 2010], learning dynamics [Erhan et al., 2010], and score matching [Vincent, 2011]. The first three aspects were already mentioned in the original paper [Vincent et al., 2008]. According to these aspects, a DAE learns one of the following: a manifold on which the data are arranged (manifold learning); the latent variables, which often behave as nonlinear coordinates in the feature space, that generate the data (generative modeling); a transformation of the data distribution that maximizes the mutual information (infomax); good initial parameters that allow the training to avoid local minima (learning dynamics); or the data distribution (score matching).

A turning point appears to be the finding of the score matching aspect [Vincent, 2011], which reveals that score matching with a special form of energy function coincides with a DAE. This means that a DAE is a density estimator of the data distribution p. In other words, it extracts and stores information as a function of p. Since then many researchers omitted stacking deterministic autoencoders, and have developed generative density estimators [Bengio et al., 2013, 2014] instead.

The generative modeling is more compatible not only with the restricted Boltzmann machine and deep belief nets [Hinton et al., 2006] and the deep Boltzmann machine [Salakhutdinov and Hinton, 2009], but also with many sophisticated algorithms, such as variational autoencoder [Kingma and Welling, 2014], minimum probability flow [Sohl-Dickstein et al., 2011, 2015], adversarial generative networks [Goodfellow et al., 2014], semi-supervised learning [Kingma et al., 2014, Rasmus et al., 2015], and image generation [Radford et al., 2016]. In generative models, what a hidden layer represents basically corresponds to either the "hidden state" itself that generates the data or the parameters (such as means and covariance matrices) of the probability distribution of the hidden states. See Bengio et al. [2014], for example.

"What do deep layers represent?" and "why deep?" are difficult questions for concrete mathematical analysis because a deep layer is a composition of nonlinear maps. In fact, even a shallow network is a universal approximator; that is, it can approximate any function, and thus, deep structure is simply redundant in theory. It has even been reported that a shallow network could outperform a deep network [Ba and Caruana, 2014]. Hence, no studies on subjects such as "integral representations of deep neural networks" or "deep ridgelet transform" exist.

Thus far, few studies have characterized the deep layer of stacked autoencoders. The only conclusion that has been drawn is the traditional belief that a combination of the "codes" exponentially enhances the expressive power of the network by constructing a hierarchy of knowledge and it is efficient to capture a complex feature of the data.

Bouvrie et al. [2009], Bruna and Mallat [2013], Patel et al. [2015] and Anselmi et al. [2015] developed sophisticated formulations for convolution networks from a group invariance viewpoint. However, their analyses are inherently restricted to the convolution structure, which is compatible with linear operators.

In this paper, we consider an autoencoder to be a transportation map and focus on its dynamics, which is a deterministic standpoint. We address the questions stated above while seeking an integral representation of a deep neural network.

2 Preliminaries

In this paper, we treat five versions of DAEs: the ordinary DAE Φ, anisotropic DAE $\Phi(\,\cdot\,; D)$, stacked DAE $h^L \circ \cdots \circ h^0$, a composition of DAEs Γ, and the continuous DAE φ.

By using the single symbols Φ and φ, we emphasize that they are realized as a shallow network or a network with a single hidden layer. Provided that there is no risk of confusion, the term "DAE Φ" without any modifiers means a shallow DAE, without distinguishing "ordinary," "anisotropic," or "continuous," because they are all derived from (3).

By ∂_t, ∇, and $\triangle$, we denote time derivative, gradient, and Laplacian, by $|\,\cdot\,|$ the Euclidean norm, by Id the identity map, and by $\mathcal{N}(\mu, \Sigma)$ the uni/multivariate Gaussian with mean μ and covariance matrix Σ.

An (anisotropic) heat kernel $\mathrm{W}_t(x, y; D)$ is the fundamental solution of an anisotropic diffusion equation on $\mathbb{R}^m$ with respect to the diffusion coefficient tensor $D(x, t)$:

$$\begin{aligned} \partial_t \mathrm{W}_t(x, y) &= \nabla \cdot [D(x, t) \nabla \mathrm{W}_t(x, y)], \quad x, y \in \mathbb{R}^m \\ \lim_{t \to 0} \mathrm{W}_t(x, y) &= \delta(x - y), \quad x, y \in \mathbb{R}^m \\ \lim_{|(x,y)| \to \infty} |\mathrm{W}_t(x, y)| &= 0, \quad t > 0. \end{aligned}$$

When $D(x, t) \equiv I$, the diffusion equation and the heat kernel are reduced to a heat equation $\partial_t \mathrm{W}_t = \triangle \mathrm{W}_t$ and a Gaussian $\mathrm{W}_t(x, y; I) := (4\pi t)^{-m/2} \exp(-|x - y|^2/4t)$. If D is clear from the context, we write simply $\mathrm{W}_t(x, y)$ without indicating D.

For a map $f : \mathbb{R}^m \to \mathbb{R}^n$ with $m \leq n$, the Jacobian $|\nabla f|$ is calculated by $\sqrt{|(\nabla f)^\top (\nabla f)|}$, regarding ∇f as an $m \times n$ matrix. By $f_\sharp p$, we denote the pushforward measure of a probability measure p with respect to a map f, which satisfies $(f_\sharp p \circ f)|\nabla f| = p$. See [Evans and Gariepy, 2015] for details.

2.1 Denoising Autoencoder

Let $X \sim p_0$ be a random vector in $M = \mathbb{R}^m$ and $\widetilde{X}$ be its corruption:

$$\widetilde{X} := X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, tI).$$

We train a shallow neural network g for minimizing an objective function

$$\mathbb{E}_{X,\widetilde{X}}|g(\widetilde{X}) - X|^2.$$

In this study, we assumed that g has a sufficiently large number of hidden units to approximate any function, and thus, the training attains the Bayes optimal. In other words, g converges to the regression function

$$\begin{aligned}\Phi(x) &:= \arg\min_g \mathbb{E}_{X,\widetilde{X}}|g(\widetilde{X}) - X|^2 \\ &= \mathbb{E}_X[X|\widetilde{X} = x],\end{aligned} \tag{2}$$

as the number of hidden units tends to infinity. We regard and treat this limit Φ as a shallow network and call it a *denoising autoencoder* or *DAE*.

Let Φ be a DAE trained for $X \sim p_0$. Denote by h and k the hidden layer and output layer of Φ, respectively; that is, they satisfy $\Phi = k \circ h$. According to custom, we call h the *encoder*, k the *decoder*, and $Z := h(X)$ the *feature* of X.

Remark on a potential confusion. Although we trained Φ as a function of $\widetilde{X}$ in order to enhance robustness, we plug in X in place of $\widetilde{X}$. Then, Φ no longer behaves as an identity map, which may be expected from traditional autoencoders, but as a denoising map formulated in (3).

2.2 Alain's Derivation of Denoising Autoencoders

Alain and Bengio [2014, Theorem 1] showed that the regression function (2) for a DAE is reduced to (1). We can rewrite it as

$$\Phi_t = \mathrm{Id} + t\nabla \log[\mathrm{W}_{t/2} * p_0], \tag{3}$$

where W_t is the isotropic heat kernel ($D \equiv I$) and p_0 is the data distribution. The proof is straightforward:

$$\begin{aligned}\frac{\mathbb{E}_\varepsilon[p_0(x-\varepsilon)(x-\varepsilon)]}{\mathbb{E}_\varepsilon[p_0(x-\varepsilon)]} &= x - \frac{\mathbb{E}_\varepsilon[p_0(x-\varepsilon)\varepsilon]}{\mathbb{E}_\varepsilon[p_0(x-\varepsilon)]} \\ &= x + \frac{t\nabla \mathrm{W}_{t/2} * p_0(x)}{\mathrm{W}_{t/2} * p_0(x)} \\ &= x + t\nabla \log[\mathrm{W}_{t/2} * p_0(x)],\end{aligned}$$

where the second equation follows by the fact that $\nabla \mathrm{W}_{t/2}(\varepsilon) = -(\varepsilon/t)\mathrm{W}_{t/2}(\varepsilon)$.

As an infinitesimal limit, (3) is reduced to an asymptotic formula:

$$\Phi_t = \mathrm{Id} + t\nabla \log p_0 + o(t^2), \quad t \to 0. \tag{4}$$

We can interpret it as a velocity field over the ground space M:

$$\partial_t \Phi_0(x) = \nabla \log p_0(x), \quad x \in M. \tag{5}$$

It implies that the initial velocity of the transportation $t \mapsto \Phi_t(x)$ of a mass on M is given by the *score*, which is in the sense of "score matching."

2.3 Anisotropic Denoising Autoencoder

We introduce the *anisotropic DAE* as

$$\Phi_t(x;D) := x + t\nabla \log \left[\int_M \mathrm{W}_t(x,y;D) p_0(y) \mathrm{d}y \right], \quad x \in M$$

by replacing the heat kernel W in (3) with an anisotropic heat kernel $\mathrm{W}(\,\cdot\,;D)$. The original formulation corresponds to the case $D(x,t) \equiv (1/2)I$.

Because of the definition, the initial velocity does not depend on $D(x,t)$. Hence, (5) still holds for the anisotropic case.

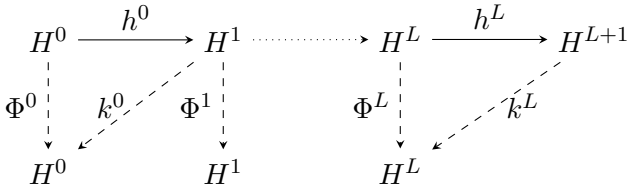
$$\partial_t \Phi_0(x;D) = \nabla \log p_0(x), \quad x \in M.$$

If D is clear from the context, we write simply $\Phi(x)$ without indicating D.

2.4 Stacked Denoising Autoencoder

Let $H^\ell\,(\ell = 0, \cdots, L+1)$ be vector spaces and Z^ℓ denote a feature vector that takes a value in H^ℓ. The input space (M) and an input vector (X) are rewritten in H^0 and Z^0, respectively. A stacked DAE is obtained by iteratively alternating (i) training a DAE $\Phi^\ell : H^\ell \to H^\ell$ for the feature Z^ℓ and (ii) extracting a new feature $Z^{\ell+1} := h^\ell(Z^\ell)$ with the encoder h^ℓ of Φ^ℓ.

We call a composition $h^L \circ \cdots \circ h^0$ of encoders a *stacked DAE*, which corresponds to the solid lines in the diagram below.



. (6)

Here, k^ℓ is the decoder of Φ^ℓ.

Remark on a potential confusion. By definition, a "stacked DAE" defined by (6) is different from a "composition of DAEs" defined by (7). While the stacked DAE is a concatenation without decoders k^ℓ, the composition of DAEs is a concatenation with decoders k_ℓ. Namely, $\Phi_L \circ \cdots \circ \Phi_0 = k_L \circ h_L \circ \cdots \circ k_0 \circ h_0$. Nevertheless, we shed light on their equivalence. See Section 3 for more details.

2.5 Composition of Denoising Autoencoders

Fix the ground space M and the data distribution p_0 on M. By $\Gamma_\tau^t : M \to M$, we symbolize a composition of DAEs $\Phi_\ell\,(\ell = 0, \cdots, L)$ on M with a fixed time interval τ.

$$M \xrightarrow{\Phi_0} M \xrightarrow{\Phi_1} \cdots \xrightarrow{\Phi_L} M \quad , \tag{7}$$

where $t := L\tau$,

$$\Phi_\ell := \mathrm{Id} + \tau \nabla \log[\mathrm{W}_{\tau/2} * p_\ell], \quad (\ell = 0, \cdots, L)$$

and

$$p_\ell := \begin{cases} p_0, & \ell = 0, \\ \Phi_{\ell\sharp} p_{\ell-1} & \ell = 1, \cdots, L. \end{cases}$$

2.6 Continuous Denoising Autoencoder

As an infinitesimal limit $\tau \to 0$ of a composition Γ_τ^t, we introduce the *continuous* DAE φ_t. See Theorem 5.1 for more details.

By the infinite compositions, we expect the relation (5) to hold for every time t. In other words, at every time t the velocity field is determined by the current score $\nabla \log p_t$. To be precise, we define $\varphi_t : M \to M$ as a flow, or the solution operator of the following continuous dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = \nabla \log p_t(x(t)), \quad t \geq 0 \tag{8}$$

where $p_t := \varphi_{t\sharp} p_0$. Hence, it satisfies

$$\varphi_0 = \mathrm{Id}, \tag{9}$$

$$\partial_t \varphi_t = \nabla \log[\varphi_{t\sharp} p_0 \circ \varphi_t], \quad t \geq 0 \tag{10}$$

which is a time-dependent continuous dynamics. Therefore, φ_t is a nonlinear semigroup:

$$\varphi_{t \to s} \circ \varphi_{0 \to t} = \varphi_{0 \to s}, \quad 0 \leq t \leq s$$

which means that a composition of continuous DAEs always coincides with a shallow DAE. Caveat: Although φ_t corresponds to having continuously deep layers, we count (a single letter) φ_t as a shallow network, as already noted at the beginning of this section.

The following integral equation is equivalent to (9) and (10)

$$\varphi_t = \mathrm{Id} + \int_0^t \nabla \log[\varphi_{s\sharp} p_0 \circ \varphi_s] \mathrm{d}s. \tag{11}$$

2.7 Ridgelet Analysis of Denoising Autoencoders

Consider the difference between a neural network g that attains the Bayes optimal (2) with a finite number of hidden units and the Bayes optimal Φ itself. While the network g has an explicit implementation

$$g(x) = \sum_{j=1}^{J} c_j \eta(a_j \cdot x - b_j), \tag{12}$$

with the activation function η, the optimal function Φ does not. Thus, the manner in which the encoder h and decoder k relate to Φ is unclear.

In this study, we regarded g as a discretization of the integral representation

$$\Phi(x) = \int \mathscr{R}\Phi(a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b, \tag{13}$$

where $\mathscr{R}\Phi(a,b)$ is the *ridgelet transform* of Φ:

$$\mathscr{R}\Phi(a,b) := \int \Phi(x)\overline{\psi(a\cdot x - b)}\mathrm{d}x,$$

with respect to an admissible ridgelet function ψ. Let $\mathscr{R}^\dagger \mathrm{T}$ denote the dual ridgelet transform of T:

$$\mathscr{R}^\dagger \mathrm{T}(x) := \iint \mathrm{T}(a,b)\eta(a\cdot x - b)\mathrm{d}a\mathrm{d}b.$$

Then, (13) is simply written as a reconstruction formula $\mathscr{R}^\dagger\mathscr{R}\Phi = \Phi$.

Although there are a number of different discretization schemes, any scheme can be symbolized by a weight function $c(a,b)$ and a probability measure $\mu(a,b)$ such that

$$\mathscr{R}\Phi(a,b) = c(a,b)\mu(a,b).$$

Then, the encoder h and decoder k correspond to a random variable $h \sim \mu$ and a weighted expectation

$$k[\,\cdot\,] = \int c(a,b)\ \cdot\ \mathrm{d}\mu(a,b).$$

In addition, the "code" of each input x corresponds to a random variable $h(x)$.

The individual implementation (12) is rewritten with an empirical measure:

$$\begin{aligned}(12) &= \int c_J(a,b)\eta(a\cdot x - b)\mathrm{d}\mu_J(a,b),\\ \mu_J(a,b) &:= \frac{1}{J}\sum_{j=1}^{J}\delta_{(a_j,b_j)},\\ c_J(a_j,b_j) &:= c_j, \quad j = 1,\cdots J,\end{aligned}$$

where we assume that $c_J(a,b)\mu_J(a,b) \to \mathscr{R}\Phi(a,b)$ as $J\to\infty$. Thus, we can understand the encoder (a vector valued function) $[\eta(a_j\cdot x - b_j)]_{j=1}^J$ and decoder (linear map) $[c_1,\cdots,c_J]$ in (12) as a collection of realizations of $h\sim\mu$ and an empirical expectation

$$\int c_J(a,b)\cdot\mathrm{d}\mu_J(a,b).$$

3 Decoding Stacked Denoising Autoencoders

Consider a stacked DAE $h^L \circ \cdots \circ h^0$ defined by (6). By using the corresponding decoders k^ℓ, we can translate it as a composition of denoising DAEs. We call this technique *decoding*. As a consequence, the investigation of stacked DAEs is reduced to the investigation of a composition of DAEs.

3.1 Decoding

Given a data distribution p_0^0 on $M_0^0 := M$, we define (M_n^ℓ, p_n^ℓ) and Φ_n^ℓ on H^ℓ trained for p_n^ℓ with diffusion coefficient D_n^ℓ by, for every n,

$$M_n^\ell := \begin{cases} h^{\ell-1}(M_{\ell-1}^{\ell-1}), & \ell = n \\ k^\ell(M_n^{\ell+1}), & 0 \le \ell < n, \end{cases}$$

$$p_n^\ell := \begin{cases} h^{\ell-1}_\sharp p_{\ell-1}^{\ell-1}, & \ell = n \\ k^\ell_\sharp p_n^{\ell+1}, & 0 \le \ell < n, \end{cases}$$

$$D_n^\ell := \begin{cases} I, & \ell = n \\ (\nabla_{H^{\ell+1}} k^\ell)^\top D_n^{\ell+1} (\nabla_{H^{\ell+1}} k^\ell), & 0 \le \ell < n. \end{cases}$$

Namely, p_ℓ^ℓ is the probability measure of the feature vector $Z^\ell \in H^\ell$, Φ_ℓ^ℓ coincides with the DAE Φ^ℓ defined in (6), and the other expressions are the (intermediate) results of decoding. By definition, each p_n^ℓ is supported in M_n^ℓ.

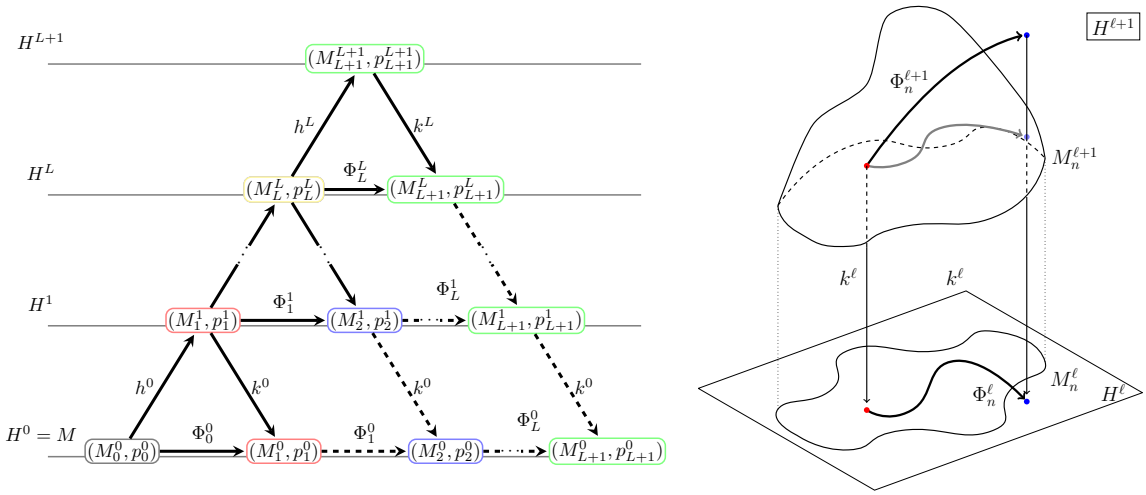


Figure 2: Diagrams of decoding. **Left**: entire decoding process. **Right**: subprocess of decoding, as an individual topological conjugation. Provided that the restrictions of decoder $k^\ell|_{M_n^{\ell+1}}$ and $k^\ell|_{M_{n+1}^{\ell+1}}$ are injective, we can pull DAE $\Phi_n^{\ell+1}$ on $H^{\ell+1}$ back into H^ℓ to obtain Φ_n^ℓ.

Figure 2 (left) depicts the decoding process, which corresponds to Theorem 3.1. The leftmost arrows correspond to the stacked DAE, the rightmost arrows to the decoders applied for decoding, the bottom arrows to the decoding result, and the inner nodes to the intermediate results of decoding, represented by (M_n^ℓ, p_n^ℓ) embedded in H^ℓ. Outlines of the same color indicate that the nodes are obtained by decoding the same feature $Z^\ell \sim p_\ell^\ell$. Every solid arrow means that it is a part of the stacked DAE, while every dashed arrow means that it is placed or obtained for decoding purposes.

Theorem 3.1. *Provided that every $h^\ell|_{M_\ell^\ell}$ is a continuous injection and every $k^\ell|_{M_n^{\ell+1}}$ is an injection, then,*

$$(k^0 \circ \cdots \circ k^L) \circ (h^L \circ \cdots \circ h^0) = \Phi_L^0 \circ \cdots \circ \Phi_0^0. \tag{14}$$

We call the application of decoders $(k^0 \circ \cdots \circ k^L)$ *decoding*. In brief, the theorem states that a stacked autoencoder followed by decoders coincides with a composition of autoencoders. In many cases, a stacked autoencoder is followed by a linear output layer k. In such a case, the decoding part $k^0 \circ \cdots \circ k^L$ is hidden in a part of the output layer as $k \circ (k^0 \circ \cdots \circ k^L)$. This implies that we can reproduce the nonlinearity of a deeper feature in the ground space.

Proof. The proof follows from multiple applications of the conjugacy $k^\ell \circ \Phi_n^{\ell+1} = \Phi_n^\ell \circ k^\ell$ explained in Lemma 3.2 and Lemma 3.3; that is,

$$\begin{aligned}
(k^0 \circ \cdots \circ k^L) \circ (h^L \circ \cdots \circ h^0) &= k^0 \circ \cdots \circ k^{L-1} \circ \Phi_L^L \circ h^{L-1} \circ \cdots \circ h^0 \\
&= k^0 \circ \cdots \circ \Phi_L^{L-1} \circ k^{L-1} \circ h^{L-1} \circ \cdots \circ h^0 \\
&= k^0 \circ \cdots \circ \Phi_L^{L-1} \circ \Phi_{L-1}^{L-1} \circ \cdots \circ h^0 \\
&\cdots \\
&= \Phi_L^0 \circ \Phi_{L-1}^0 \circ \cdots \circ \Phi_0^0.
\end{aligned}$$

As the proof suggests, decoding implicitly *factorizes* $H^\ell \to H^{\ell+1}$ into $H^\ell \to M \to H^{\ell+1}$, as depicted in Figure 1. In the case of a stacked continuous DAE, the RHS is reduced to a shallow DAE φ_t.

3.2 Topological Conjugacy

For decoding, we insist on using k^ℓ in common, because it is *easy to implement* as shown in Figure 3 and Figure 4. Clearly, it is *not* obvious that we can use k^ℓ in common for decoding the entire autoencoder, because k^ℓ is simply defined as a part of Φ_ℓ^ℓ. Lemma 3.2 and Lemma 3.3 guarantees that k^ℓ works.

Figure 2 (right) depicts an individual subprocess of decoding, which corresponds to Lemma 3.2 and Lemma 3.3. Although each H^ℓ may differ in its dimensions, every M_n^ℓ has *the same* dimension as the ground space M, provided that every $h^\ell|_{M_\ell^\ell}$ and $k^\ell|_{M_n^{\ell+1}}$ is homeomorphic. In such a case, we can pull $\Phi_n^{\ell+1}$ on $H^{\ell+1}$ back to H^ℓ. Lemma 3.2 and Lemma 3.3 state that the pullback coincides with Φ_n^ℓ, and they are *topological conjugate* to each other. See Appendix A for the proof.

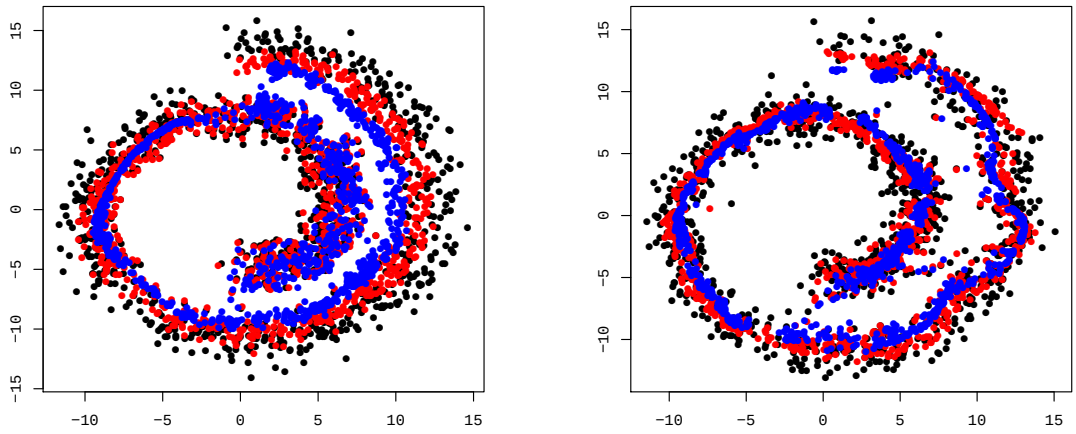


Figure 3: Activation patterns of two different networks trained for two-dimensional Swiss roll data X (**black**). **Left**: A stacked denoising autoencoder $h^1 \circ h^0$ visualized by decoding $k^0 \circ h^0(X)$ (**red**) and $(k^0 \circ k^1) \circ (h^1 \circ h^0)(X)$ (**blue**). **Right**: A composition of denoising autoencoders $\Phi_1 \circ \Phi_0$ simply visualized by $\Phi_0(X)$ (**red**) and $\Phi_1 \circ \Phi_0(X)$ (**blue**). Although the two networks are trained by different algorithms, they resemble each other.

Lemma 3.2. *Let $\Phi_n^{\ell+1}$ and Φ_n^{ℓ} be anisotropic DAEs defined in the diagram. Assume that the domain and range of $\Phi_n^{\ell+1}$ are contained in $M_n^{\ell+1}$ and $M_{n+1}^{\ell+1}$, respectively, and the restrictions $k^\ell|_{M_n^{\ell+1}}$, $k^\ell|_{M_{n+1}^{\ell+1}}$ and $\Phi_n^{\ell+1}|_{M_n^{\ell+1}}$ are injective, respectively. Then, $\Phi_n^{\ell+1}$ and Φ_n^{ℓ} are topologically conjugate to each other with a conjugation map k^ℓ. That is,*

$$k^\ell \circ \Phi_n^{\ell+1} = \Phi_n^{\ell} \circ k^\ell. \tag{15}$$

In particular, $\Phi_n^{\ell}|_{M_n^{\ell}}$ is an injection toward M_{n+1}^{ℓ}.

By definition, the inverse $(k^\ell|_{M_n^{\ell+1}})^{-1}$ exists, although it is not a linear map. Thus, (15) implies the topological conjugacy.

A similar statement holds for continuous DAEs. Contrary to ordinary DAEs, the pullback of a continuous DAE is independent of k^ℓ. This is mainly because a continuous DAE is ruled by dynamics at every time and there is no room for developing the variety of encoders h^ℓ.

Lemma 3.3. *Let $\varphi_t^{\ell+1}$ and φ_t^{ℓ} be continuous DAEs trained for $p_n^{\ell+1}$ and p_n^{ℓ}, respectively. Assume that the domain and range of $\varphi_t^{\ell+1}$ are contained in $M_n^{\ell+1}$ and $M_{n+1}^{\ell+1}$, respectively, and the restrictions $k^\ell|_{M_n^{\ell+1}}$, $k^\ell|_{M_{n+1}^{\ell+1}}$ and $\varphi_n^{\ell+1}|_{M_n^{\ell+1}}$ are injective, respectively. Then, $\varphi_t^{\ell+1}$ and φ_t^{ℓ} are topologically conjugate to each other with a conjugation map k^ℓ;*

that is,

$$k^\ell \circ \varphi_t^{\ell+1} = \varphi_t^\ell \circ k^\ell. \tag{16}$$

In particular, $\varphi_t^\ell|_{M_n^\ell}$ *is an injection toward* M_{n+1}^ℓ.

3.3 Integral Representation of Stacked Denoising Autoencoder

Recall that a neural network g that approximates $\Phi : M \to M$ with a *single* hidden layer is obtained by discretizing the integral representation $\mathscr{R}^\dagger\mathscr{R}\Phi$. As a consequence of Theorem 3.1, we can define the *integral representation of a stacked DAE* as

$$\begin{aligned}
&(k^0 \circ \cdots \circ k^L) \circ (h^L \circ \cdots \circ h^0) \\
&\quad = \mathscr{R}^\dagger\mathscr{R}\Phi_L^0 \circ \cdots \circ \mathscr{R}^\dagger\mathscr{R}\Phi_0^0 \\
&\quad = \int \mathscr{R}\Phi_L^0(a_L, b_L)\eta \left[a_L \cdots \left(\int \mathscr{R}\Phi_0^0(a_0, b_0)\eta(a_0 \cdot - b_0)\mathrm{d}a_0\mathrm{d}b_0 \right) \cdots - b_L \right] \mathrm{d}a_L\mathrm{d}b_L.
\end{aligned} \tag{17}$$

In the case of a continuous DAE, we can *compress* the deep network in RHS to a shallow network:

$$\mathscr{R}^\dagger\mathscr{R}\varphi^0 = \mathscr{R}^\dagger\mathscr{R}\varphi_L^0 \circ \cdots \circ \mathscr{R}^\dagger\mathscr{R}\varphi_0^0. \tag{18}$$

Let us consider $h^1 \circ h^0$ for example. The integral representation is calculated as

$$\begin{aligned}
&\int \mathscr{R}\Phi_1^0(a_1, b_1)\eta \left(a_1 \cdot \int \mathscr{R}\Phi_0^0(a_0, b_0)\eta(a_0 \cdot x - b_0)\mathrm{d}a_0\mathrm{d}b_0 - b_1 \right) \mathrm{d}a_1\mathrm{d}b_1 \\
&= \int c_1(a_1, b_1)\eta \left(\int a_1 \cdot c_0(a_0, b_0)\eta(a_0 \cdot x - b_0)\mathrm{d}\mu_0(a_0, b_0) - b_1 \right) \mathrm{d}\mu_1(a_1, b_1) \\
&\approx \sum_i c_1^{(i)}\eta \left(\sum_j a_1^{(i)} \cdot c_0^{(j)}\eta \left(a_0^{(j)} \cdot x - b_0^{(j)} \right) - b_1^{(i)} \right),
\end{aligned}$$

where we defined $\mathscr{R}\Phi_\ell^0(a_\ell, b_\ell) =: c_\ell(a_\ell, b_\ell)\mu_\ell(a_\ell, b_\ell)$ for $\ell = 0, 1$. The last equation is numerically tractable and hence indicates what a stacked layer represents in a most direct form. Observe that the second hidden coefficients are factorized as $a_1 \cdot c_0(a_0, b_0)$. This means that decoding reduced the multiple application $\mathscr{R}|_{H^1} \circ \mathscr{R}|_{H^0}$, which accompanies the ridgelet transform on a stacked layer, to a composition $\mathscr{R}|_{H^0} \circ \mathscr{R}^\dagger\mathscr{R}|_{H^0}$.

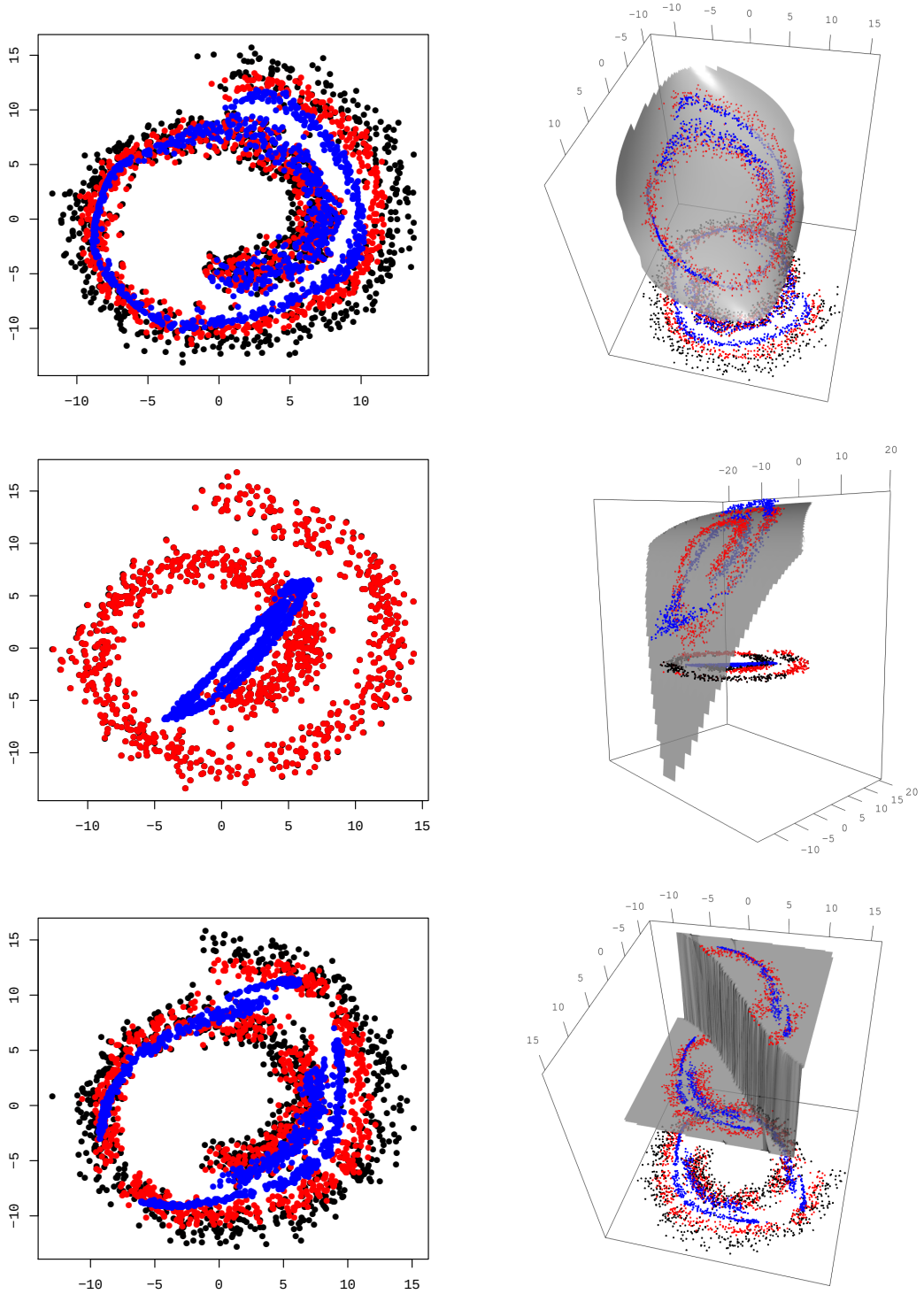


Figure 4: Decoding results of stacked denoising autoencoders. The **right** side corresponds to Figure 2 (right); the **gray surface**, **red** points, and **blue** points correspond to M_1^1 embedded in H^1, $h^0(X)$, and $k^1 \circ h^1 \circ h^0(X)$, respectively. **Top**: M_1^1 is nearly flat and the decoding results are almost isotropic. **Middle**: M_1^1 is highly curved and the decoding results are anisotropic. **Bottom**: M_1^1 is separated and the decoding results are also separated.

4 Continuous Denoising Autoencoder Solves Backward Heat Equation

Instead of directly analyzing a composition of anisotropic DAEs, we focus on the continuous DAE. The following provides suggestions for obtaining and understanding continuous DAEs.

Theorem 4.1. *Let p_0 be a data distribution on $M = M$ and $\varphi_t : M \to M$ be the continuous DAE trained for p_0. Then, the pushforward measure $p_t := \varphi_{t\sharp} p_0$ satisfies the* backward heat equation*:*

$$\partial_t p_t = -\triangle p_t, \quad p_{t=0} = p_0. \tag{19}$$

Observe that the backward heat equation (19) is equivalent to the following *final value problem* for the ordinary heat equation:

$$\partial_t u_t = \triangle u_t, \quad u_{t=\mathrm{T}} = p_0 \quad \text{for some } \mathrm{T}$$

where u_t denotes a probability measure on M. Indeed,

$$p_t = u_{\mathrm{T}-t},$$

is the solution of (19). See Appendix B for more details.

Thus, we can solve (19) through the ordinary heat equation. When we have obtained p_{T}, we can obtain p_t by

$$p_t = \mathrm{W}_{\mathrm{T}-t} * p_{\mathrm{T}}, \quad 0 \le t \le \mathrm{T}.$$

For example, we can combine it with (11) and obtain another integral equation:

$$\varphi_t(x) = x + \int_0^t \nabla \log[\mathrm{W}_{\mathrm{T}-s} * p_{\mathrm{T}}] \circ \varphi_s(x) \mathrm{d}s.$$

According to Wasserstein geometry, or the geometric aspect of transportation theory, the heat equation is the *(abstract) gradient flow* that increases the Boltzmann entropy functional $H[p] := \mathbb{E}_p[-\log p]$ [Villani, 2009, Th. 23.19]. Here the term "abstract" emphasizes that the gradient and flow are defined on the space of probability measures equipped with a Wasserstein metric, or the Wasserstein space. As a consequence, we can understand the continuous DAE as an abstract gradient flow that decreases the entropy.

$$\partial_t p_t = -\nabla H[p_t],$$

where ∇ is defined in the weak sense, or the sense of Otto's analysis [Otto and Villani, 2000]. This reformulation promotes our understanding, because this system does *not* depend on time. See Figure 5, for example. Brenier's theorem [Brenier, 1991] guarantees that such a flow φ_t exists and it is a potential flow.

In general, the estimation of the time reversal of the diffusion process is an inverse problem. For example, when we solve the backward heat equation by a difference method, it is unstable in the sense of L^∞-norm, and we should employ some regularization techniques. The investigation of stability is our important future work.

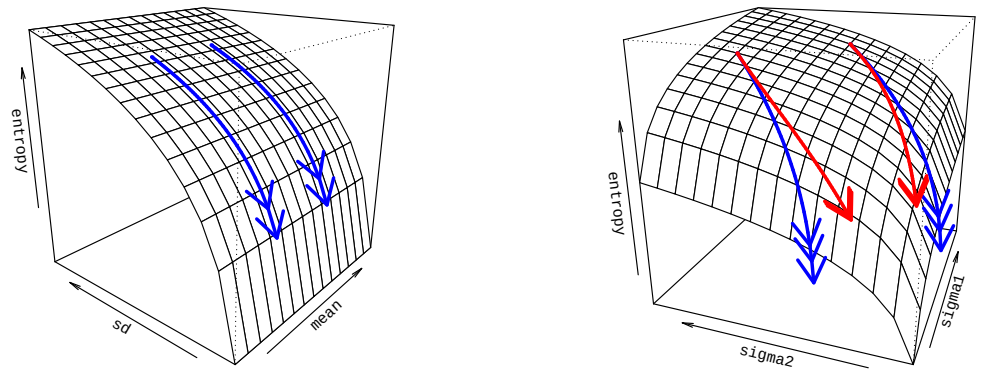


Figure 5: Abstract orbits of pushforward measures in the space of probability measures. Each surface corresponds to an entropy functional. **Left**: the space of univariate Gaussians $\mathcal{N}(\mu, \sigma^2)$ with entropy $(1/2)\log\sigma^2 + (\text{const.})$. **Right**: the space of bivariate Gaussians $\mathcal{N}([0,0], \mathrm{diag}\,[\sigma_1^2, \sigma_2^2])$ with entropy $(1/2)\log|\mathrm{diag}\,[\sigma_1^2, \sigma_2^2]| + (\text{const.})$. A pushforward measure $t \mapsto \varphi_{t\sharp}p_0$ (**blue**) of a continuous denoising autoencoder coincides with an abstract gradient flow with respect to entropy. A pushforward measure $(t \mapsto)\Phi_{t\sharp}p_0$ (**red**) of an ordinary denoising autoencoder gradually leaves the flow.

5 Comparison of Continuous Denoising Autoencoder and Ordinary Denoising Autoencoders

As an idealized model of a composition of DAEs Γ_τ^t, we investigate a continuous DAE φ_t, which is analytically more tractable and richer in geometric properties. Here, we justify the continuous DAE from three aspects.

5.1 Infinitesimal Composition Results in a Continuous Denoising Autoencoder

Given an initial point $x_0 \in M$, the orbit $t \mapsto \Gamma_\tau^t(x_0)$ corresponds to a Eulerian finite difference approximation, or a piecewise linear approximation of $t \mapsto \varphi_t(x_0)$. The following is fundamental for justifying our understanding.

Theorem 5.1. *Let Γ_τ^t be a composition of DAEs and φ_t be a continuous DAE trained for p_0. Provided that there exists an open domain Ω such that $\log p_0$ is Lipschitz continuous in Ω, then for every $x \in \Omega$,*

$$\lim_{\tau \to 0} \Gamma_\tau^t(x) = \varphi_{0 \to t}(x). \tag{20}$$

Sketch of Proof. Recall that the initial velocity $\partial_t \Phi_0(\,\cdot\,; D)$ does not depend on the diffusion coefficient D; then, we can assume that a Γ_τ^t is composed of ordinary isotropic DAEs without loss of generality. If $\log p_0$ is Lipschitz continuous, we can upper bound the difference $|\Gamma_\tau^t(x) - \varphi_{0\to t}(x)|$ by τ. Then, it converges to zero when $\tau \to 0$. The limit function Γ_τ^t satisfies (10) at every t. Hence, by the uniqueness of the solution, it is a continuous DAE.

5.2 Denoising Autoencoder Performs Mean-Shift

For a *well-separated* mixture of Gaussians, both ordinary and continuous DAEs perform a similar function. Namely, they transport mass toward the corresponding cluster center. Hence, we can identify a DAE as mean-shift (or k-means, EM algorithm for GMM, or some other kernel density estimator.)

Here, we consider a mixture of Gaussians $\sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k)$ is *well-separated*, when for every cluster center μ_k there exists a neighborhood Ω_k of μ_k such that $\mathcal{N}(\Omega_k; \mu_k, \Sigma_k) \approx 1$ and $\gamma_{kt} \approx \mathbf{1}_{\Omega_k}$, where γ_{kt} is the *responsibility function* defined in (42) for a continuous DAE and (45) for a DAE, respectively. Intuitively, this assumption implies that every component overlaps less and scatters, and hence, they are separated from each other. The following is an immediate consequence of Appendix F.1 and Appendix F.2.

Proposition 5.2. *Let $p_0 := \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k)$ be a well-separated mixture of Gaussians. Then, φ_t and Φ_t push p_0 forward to*

$$\varphi_{t\sharp} p_0 = \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k - 2tI),$$
$$\Phi_{t\sharp} p_0 \approx \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k (I + t\Sigma_k^{-1})^{-2}).$$

Under a well-separated condition, the restrictions $\varphi_{tk} := \varphi_t\big|_{\Omega_k}$ and $\Phi_{tk} := \Phi_t\big|_{\Omega_k}$ are contractions, because they each have the fix point μ_k and the covariance Σ_k shrinks as t increases. Hence, a DAE transports each point x toward the corresponding clustering center.

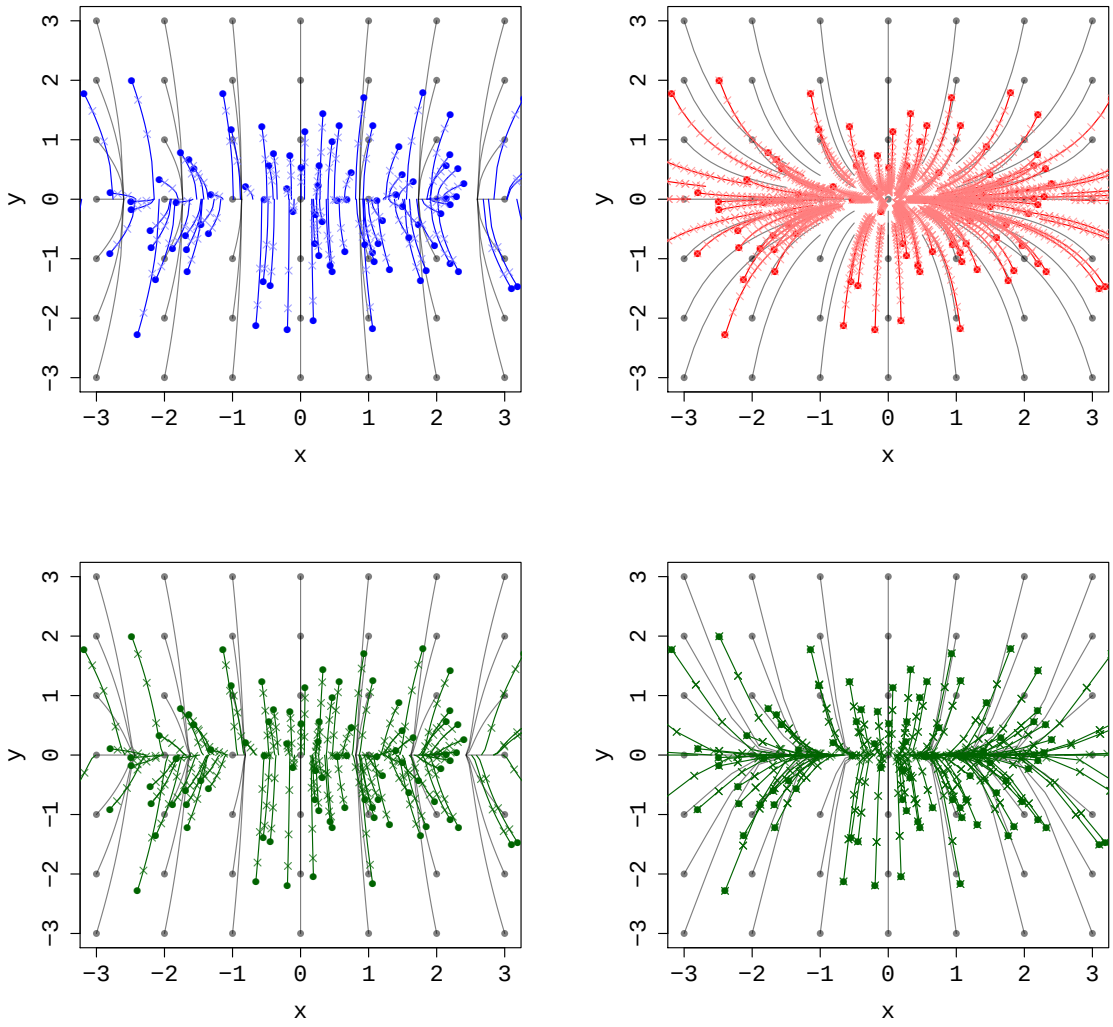


Figure 6: Orbits of denoising autoencoders trained for the same data distribution $p_0 = \mathcal{N}([0,0], \mathrm{diag}\,[2,1])$. The orbits are analytically calculated by using (36) and (38). **Top Left**: continuous denoising autoencoder $t \mapsto \varphi_t$. **Top Right**: ordinary denoising autoencoder $t \mapsto \Phi_t$. **Bottom Left**: compositions of denoising autoencoders $t \mapsto \Gamma_\tau^t$ with $\tau = 0.05$. **Bottom Right**: $\tau = 0.5$. **Gray lines** start from the regular grid. **Colored lines** start from the samples drawn from p_0. **Midpoints** are plotted every $\tau = 0.2$. The continuous denoising autoencoder attains the singularity at $t = 1/2$. The ordinary denoising autoencoder slows down as $t \to \infty$ and never attains the singularity in finite time. As τ tends to zero, Γ_τ^t draws a similar orbit to the continuous denoising autoencoder φ_t. Even the curvature of orbits changes according to τ.

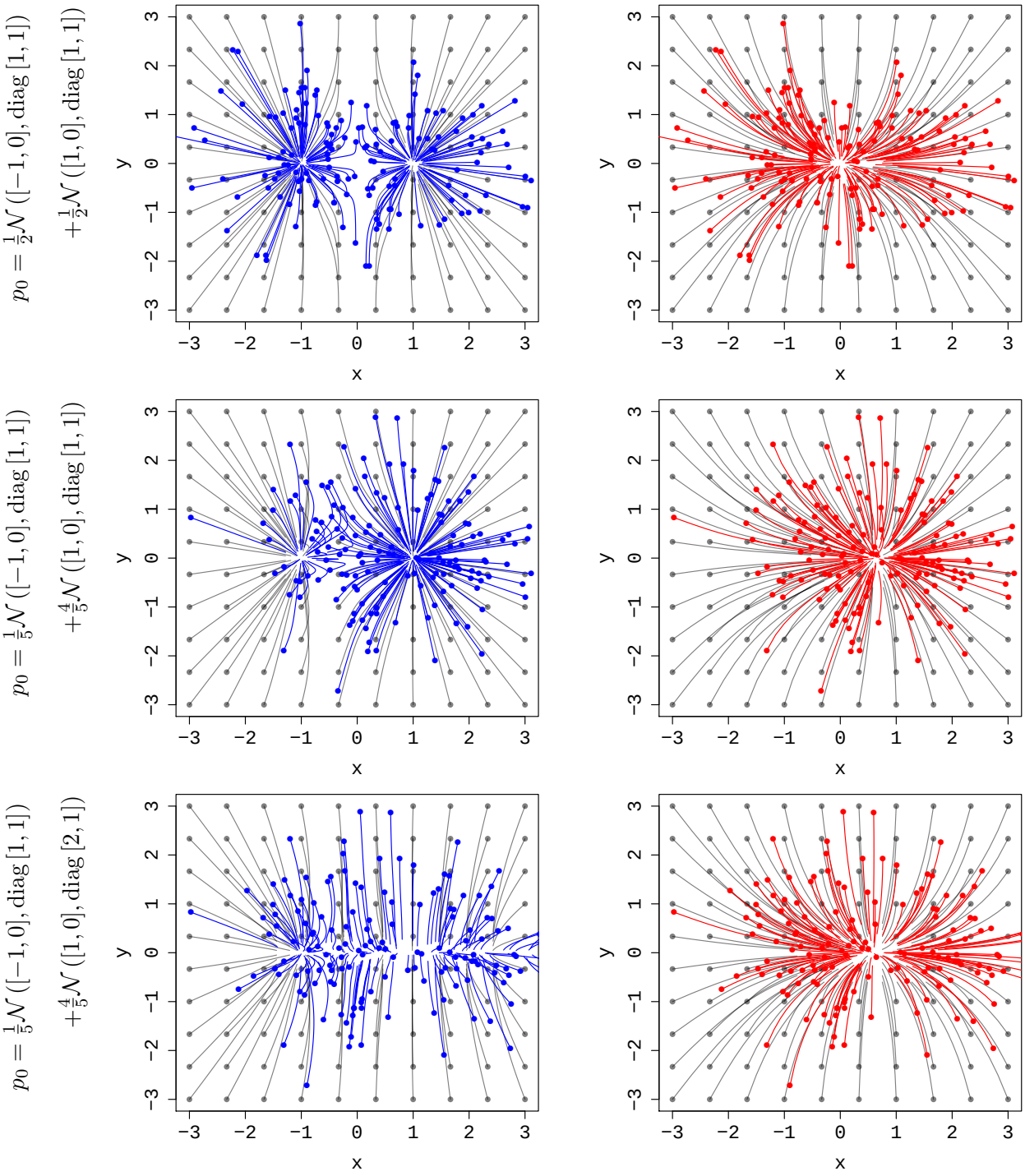


Figure 7: Orbits of continuous denoising autoencoders $t \mapsto \varphi_t$ (**left**) and ordinary denoising autoencoders $t \mapsto \Phi_t$ (**right**) trained for GMMs. The orbits are obtained by numerically solving (8) for φ_t, and by using (43) for Φ_t. At first, both denoising autoencoders transport mass toward the cluster centers. **Right**: all mass concentrates on the origin as $t \to \infty$. **Top Left**: the orbits split. **Middle Left**: some orbits meet, which reflects the time-dependency of the velocity field. **Bottom Left**: If the covariance is not homogeneous, then the mass concentrates on a manifold.

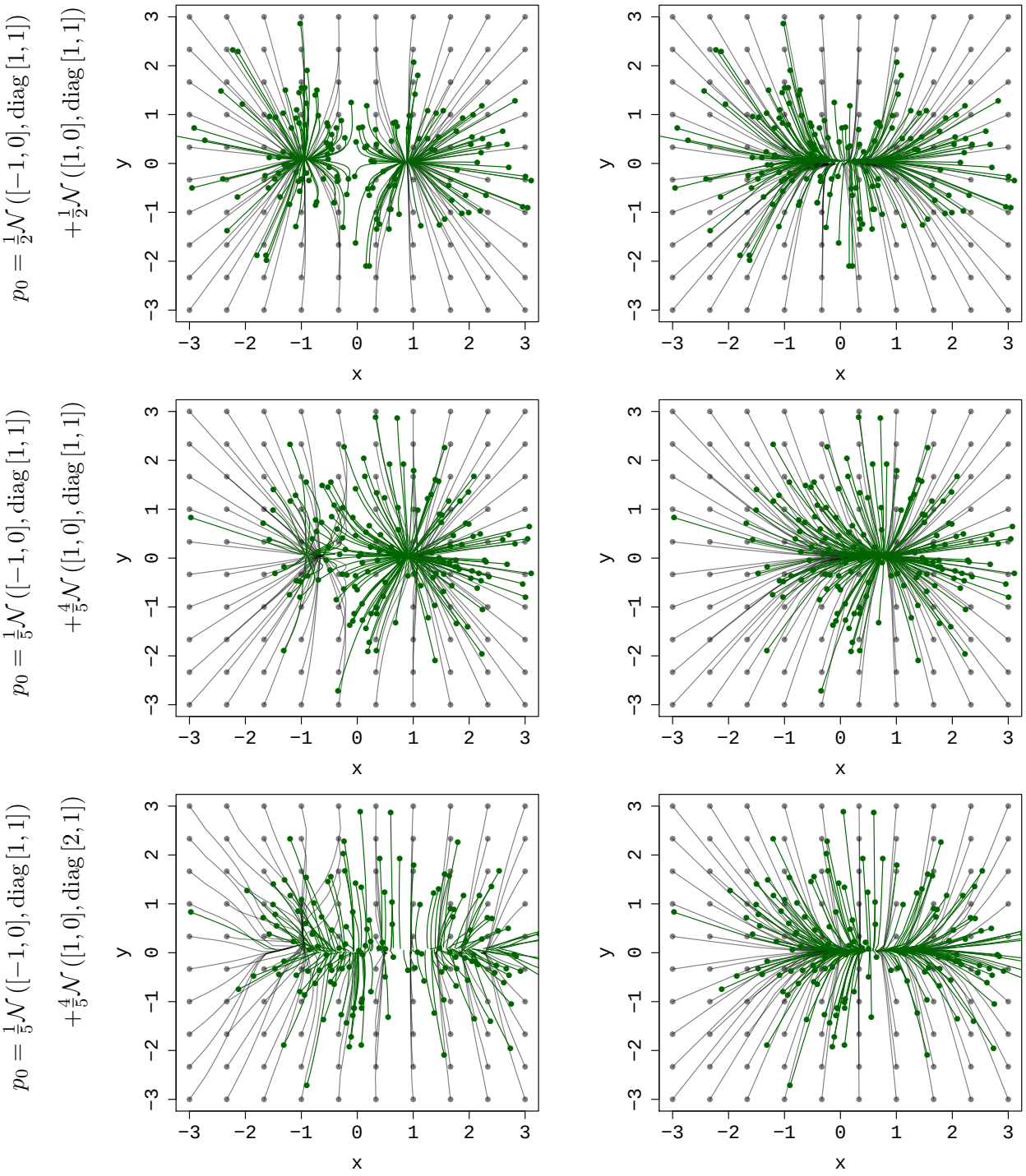


Figure 8: Orbits of compositions of denoising autoencoders $t \mapsto \Gamma_\tau^t$ with $\tau = 0.05$ (**left**) and $\tau = 0.5$ (**right**). Each Γ_τ^t is trained for the same p_0 at the same place in Figure 7. The orbits are calculated by repeatedly using (43) and estimating p_t. For the same p_0, the left side is similar to φ_t and the right side is similar to Φ_t. The smaller τ becomes, or the more frequently the number of hidden layers increases, the more similar to Φ_t Γ_τ^t becomes.

5.3 Composition Accelerates the Convergence

According to Proposition 5.2, $\varphi_{t\sharp}$ converges in *finite* time. For example, for a univariate Gaussian $\mathcal{N}(\mu, \sigma_0^2)$, it converges to Dirac's δ_μ at $t = \sigma_0^2/2$. On the other hand, $\Phi_{t\sharp}$ converges as slowly as *reciprocal* $o(t^{-2})$. Proposition 5.3 states that the composition of DAEs *accelerates* the convergence as fast as *exponential* decay. See Appendix C for the proof.

Proposition 5.3. *Let p_0 be a well-separated mixture of Gaussians and Γ_τ^t be a composition of denoising autoencoders trained for p_0. The pushforward $\Gamma_{\tau\sharp}^t p_0$ with fixed τ converges with exponential decay $o(e^{-t})$, and $\lim_{\tau\to 0} \Gamma_{\tau\sharp}^t p_0 = \varphi_{(0\to t)\sharp} p_0$.*

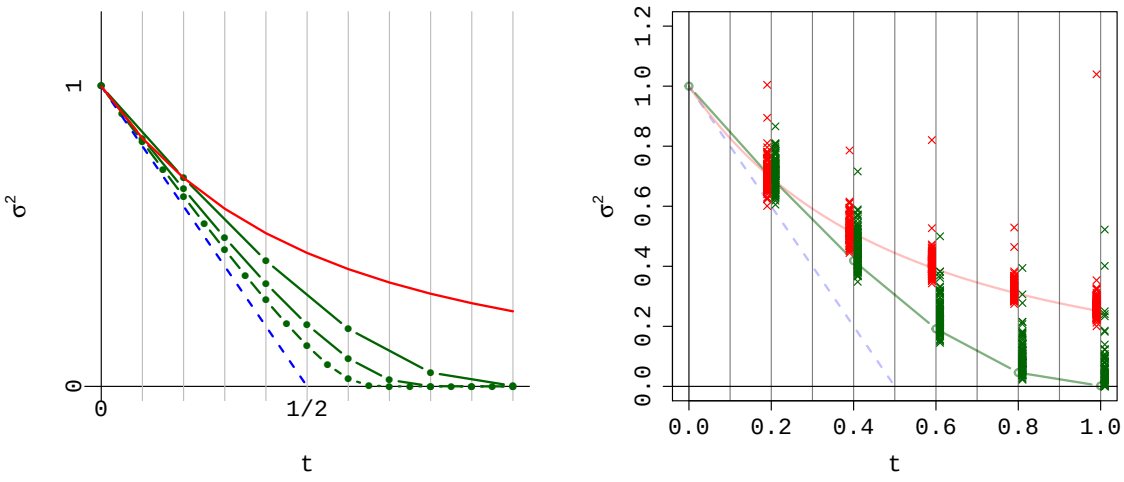


Figure 9: Time evolution of the variance σ_t^2 of a univariate Gaussian $p_0 = \mathcal{N}(0, 1)$. **Left**: analytical calculation results according to (33) and (35). A continuous denoising autoencoder $\varphi_{t\sharp} p_0$ (**blue, solid**), compositions of denoising autoencoders $\Gamma_{\tau\sharp}^t p_0$ (**green, dash-dotted**) with $\tau = 0.2, 0.1$ and 0.05 (from **up** to **down**), and an ordinary denoising autoencoder $\Phi_{t\sharp} p_0$ (**red, solid**). The variances shrink according to reciprocal $o(1/t)$ by the denoising autoencoder, exponential $o(e^{-t})$ by composition of denoising autoencoders, and linear $o(t)$ by continuous denoising autoencoder. **Right**: training results. A composition of denoising autoencoders $\Gamma_{\tau\sharp}^t p_0$ with $\tau = 0.2$ (**green**), and an ordinary denoising autoencoder $\Phi_{t\sharp} p_0$ (**red**).

6 Discussions

We investigated stacked DAEs from the viewpoints of transportation theory and ridgelet analysis.

We developed *decoding*, which is a simple technique to translate a stacked DAE into a composition of DAEs (Theorem 3.1). This follows from Lemma 3.2 and Lemma 3.3, which state that the pullback of a DAE on a stacked layer into the base layer coincides with a DAE on the base layer. As a result, we derived an *integral representation of a stacked DAE* (17), which is the first achievement in ridgelet analysis.

We introduced the *continuous DAE*, the velocity field of which is determined by the score of the current data distribution. We found that it solves the *backward heat equation* (Theorem 4.1). By definition, it is powerful in analytic properties, such as semigroup, integral equation, and heat kernel. In addition, it is rich in geometric interpretation, such as velocity field as score, backward heat equation, and abstract gradient flow with respect to entropy. It is our important future work to characterize other feature extractors for classification by some potential functional over the space of data distributions, since the classification difficulty heavily depends on the arrangements of individual observations.

In our opinion, the continuous DAE is an appropriate model of the composition of ordinary DAEs at least in three aspects. First, a composition of DAEs is a piecewise linear approximation of a continuous DAE and both coincide with each other in the limit (Theorem 5.1). Second, both continuous and ordinary DAEs perform mean-shift for GMM; that is, they transport mass toward the centers of each component Gaussians (Proposition 5.2). Third, the composition accelerates mean-shift as fast as exponential decay and again it converges to a continuous DAE in the limit (Proposition 5.3).

What do deep layers represent? A discretization of the integral representation (17) is the most straightforward answer to this question. The integral representation is obtained via ridgelet transform of a stacked DAE. By decoding, it is equivalent to a composition of DAEs, which transports mass toward the cluster centers so as to decreased entropy.

Why deep? The answer to this question is (i) deeper DAEs can perform better in extracting a "manifold" in the sense of manifold learning, and (ii) the depth accelerates the feature extraction as fast as exponential decay. See Figure 7 for example. A continuous DAE φ_t corresponds to an infinitely deep network, while an ordinary DAE Φ_t is a shallow network. If the covariance of the data distribution is not homogeneous, then φ_t transports mass toward a manifold. In addition, as Proposition 5.3 suggests, the depth accelerates the feature extraction. On the other hand, Φ_t eventually transforms all mass to the origin as $t \to \infty$, which is inevitable because of the term $\mathrm{W}_{t/2} * p_0$.

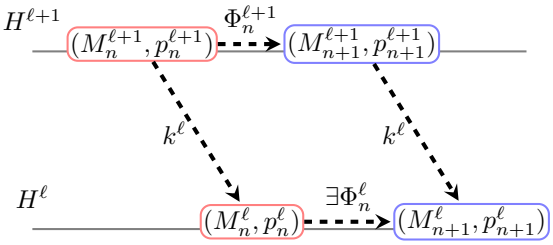
Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 15J07517.

A Proof that Stacking is Composing

A.1 Proof of Lemma 3.2

We show that the following diagram commutes. Here, $\Phi_n^{\ell+1}$ and Φ_n^{ℓ} are the anisotropic



DAEs trained for $p_n^{\ell+1}$ and p_n^{ℓ} with diffusion coefficients $D_n^{\ell+1}$ and D_n^{ℓ} and are expressed as

$$\Phi_n^{\ell+1} = \mathrm{Id}_{H^{\ell+1}} + \tau \nabla_{H^{\ell+1}} \log \left[\int_{H^{\ell+1}} \mathrm{W}_\tau(\,\cdot\,, w; D_n^{\ell+1}) p_n^{\ell+1}(w) \mathrm{d}w \right], \tag{21}$$

$$\Phi_n^{\ell} = \mathrm{Id}_{H^{\ell}} + \tau \nabla_{H^{\ell}} \log \left[\int_{H^{\ell}} \mathrm{W}_\tau(\,\cdot\,, y; D_n^{\ell}) p_n^{\ell}(y) \mathrm{d}y \right]. \tag{22}$$

Proof. By the assumption that $k^\ell|_{M_n^{\ell+1}}$ is injective, there exists a section, or the right inverse of $k^\ell : H^{\ell+1} \to H^\ell$ defined by

$$g_n := (k^\ell|_{M_n^{\ell+1}})^{-1} : M_n^\ell \to M_n^{\ell+1} \subset H^{\ell+1}. \tag{23}$$

By the assumption that $\Phi_n^{\ell+1}$ is supported in $M_n^{\ell+1}$, we can identify $\Phi_n^{\ell+1}$ as a restriction $\Phi_n^{\ell+1}|_{M_n^{\ell+1}}$. Hence, we can pull $\Phi_n^{\ell+1}|_{M_n^{\ell+1}}$ back to H^ℓ by $k^\ell : H^{\ell+1} \to H^\ell$.

According to (21) and (22),

$$k^\ell \circ \Phi_n^{\ell+1} = k^\ell + \tau \nabla_{H^\ell} \log \left[\int_{H^{\ell+1}} \mathrm{W}_\tau(\,\cdot\,, w; D_n^{\ell+1}) p_n^{\ell+1}(w) \mathrm{d}w \right], \tag{24}$$

$$\Phi_n^{\ell} \circ k^\ell = k^\ell + \tau \nabla_{H^\ell} \log \left[\int_{H^{\ell}} \mathrm{W}_\tau(k^\ell(\,\cdot\,), y; D_n^{\ell}) p_n^{\ell}(y) \mathrm{d}y \right]. \tag{25}$$

Recall that $p_n^{\ell+1}$ and p_n^{ℓ} are supported in $M_n^{\ell+1}$ and M_n^{ℓ}, respectively. Hence, by changing a variable $w \leftarrow g_n(y)$,

$$\begin{aligned} \int_{H^{\ell+1}} \mathrm{W}_\tau(\,\cdot\,, w; D_n^{\ell+1}) p_n^{\ell+1}(w) \mathrm{d}w &= \int_{M_n^{\ell+1}} \mathrm{W}_\tau(\,\cdot\,, w; D_n^{\ell+1}) p_n^{\ell+1}(w) \mathrm{d}w|_{M_n^{\ell+1}} \\ &= \int_{M_n^{\ell}} \mathrm{W}_\tau(\,\cdot\,, g_n(y); D_n^{\ell+1}) p_n^{\ell+1} \circ g_n(y) |\nabla g_n(y)| \mathrm{d}y \\ &= \int_{H_n^{\ell}} \mathrm{W}_\tau(\,\cdot\,, g_n(y); D_n^{\ell+1}) p_n^{\ell}(y) \mathrm{d}y, \end{aligned} \tag{26}$$

where $\mathrm{d}w|_{M_n^{\ell+1}}$ is the Hausdorff measure on $M_n^{\ell+1}$ and the third equation follows by the definition of the pushforward measure.

We now show that $\mathrm{W}_\tau(g_n(x), g_n(y); D_n^{\ell+1})$ satisfies the heat equation of $\mathrm{W}_\tau(x, y; D_n^\ell)$. According to the definition of the anisotropic heat kernel,

$$\begin{aligned}
\partial_t \mathrm{W}_\tau(g_n(x), g_n(y); D_n^{\ell+1}) &= \nabla_z \cdot \left[D_n^{\ell+1} \nabla_z \mathrm{W}_\tau(z, g_n(y); D_n^{\ell+1}) \right] \Big|_{z=g_n(x)} \\
&= \nabla_x \cdot \left[(\nabla_x k^\ell)^\top D_n^{\ell+1} (\nabla_x k^\ell) \nabla_x \mathrm{W}_\tau(g_n(x), g_n(y); D_n^{\ell+1}) \right] \\
&= \nabla_x \cdot \left[D_n^\ell \nabla_x \mathrm{W}_\tau(g_n(x), g_n(y); D_n^{\ell+1}) \right],
\end{aligned}$$

where the last equation follows by pulling back the gradient $(\nabla_z f) \circ g_n = (\nabla_z g_n^{-1})^\top \nabla_x (f \circ g_n)$. Hence, by the uniqueness of the solution,

$$\mathrm{W}_\tau(g_n(x), g_n(y); D_n^{\ell+1}) = \mathrm{W}_\tau(x, y; D_n^\ell). \tag{27}$$

By combining (27) and (26) with (24), we have (24) = (25).

A.2 Proof of Lemma 3.3

Let $\varphi_t^{\ell+1}$ and φ_t^ℓ be continuous DAEs put in place of $\Phi_n^{\ell+1}$ and Φ_n^ℓ in the diagram; that is, they satisfy

$$\partial_t \varphi_t^{\ell+1} = \nabla_{H^{\ell+1}} \log \left[p_t^{\ell+1} \circ \varphi_t^{\ell+1} \right], \tag{28}$$

$$\partial_t \varphi_t^\ell = \nabla_{H^\ell} \log \left[p_t^\ell \circ \varphi_t^\ell \right], \tag{29}$$

where $p_t^{\ell+1} := \varphi_{t\sharp}^{\ell+1} p_n^{\ell+1}$ and $p_t^\ell := \varphi_{t\sharp}^\ell p_n^\ell$.

Take g_n defined in (23). We show that $\widetilde{\varphi_t^{\ell+1}} := k^\ell \circ \varphi_t^{\ell+1} \circ g_n$ satisfies (29).

Proof. By differentiating $k^\ell \circ \varphi_t^{\ell+1} \circ g_n$ by t,

$$\begin{aligned}
\partial_t \widetilde{\varphi_t^{\ell+1}} &= k^\ell \circ \partial_t (\varphi_t^{\ell+1} \circ g_n) \\
&= \nabla_{H^\ell} \log \left[p_t^{\ell+1} \circ \varphi_t^{\ell+1} \circ g_n \right] \\
&= \nabla_{H^\ell} \log \left[p_n^{\ell+1} \circ g_n \left| \nabla_{H^{\ell+1}} \varphi_t^{\ell+1} \circ g_n \right|^{-1} \right] \\
&= \nabla_{H^\ell} \log \left[p_n^\ell \left| \nabla_{H^{\ell+1}} k^\ell \right| \left| \nabla_{H^{\ell+1}} \varphi_t^{\ell+1} \circ g_n \right|^{-1} \right] \\
&= \nabla_{H^\ell} \log \left[p_n^\ell \left| \nabla_{H^\ell} \widetilde{\varphi_t^{\ell+1}} \right|^{-1} \right] \\
&= \nabla_{H^\ell} \log \left[p_t^\ell \circ \widetilde{\varphi_t^{\ell+1}} \right],
\end{aligned}$$

where the third, fourth, and sixth equations follow by the definitions of pushforward:

$$\begin{aligned} p_t^{\ell+1} \circ \varphi_t^{\ell+1} \left| \nabla_{H^{\ell+1}} \varphi_t^{\ell+1} \right| &= p_n^{\ell+1}, \\ p_n^{\ell} \circ k^{\ell} \left| \nabla_{H^{\ell+1}} k^{\ell} \right| &= p_n^{\ell+1}, \\ p_t^{\ell} \circ \widetilde{\varphi_t^{\ell+1}} \left| \nabla_{H^{\ell}} \widetilde{\varphi_t^{\ell+1}} \right| &= p_n^{\ell}. \end{aligned}$$

The last equation concludes that $\widetilde{\varphi_t^{\ell+1}}$ satisfies (29). Hence, by the uniqueness of the solution, we have $k^{\ell} \circ \varphi_t^{\ell+1} = \varphi_t^{\ell} \circ k^{\ell}$.

B Proof of Theorem 4.1

Fix the initial time $t_0 = 0$ and an input distribution $p_0(X)$. Let Φ_t be the DAE trained for $p_0(X)$ and write the pushforward measure as $p_t := \Phi_{t\sharp} p_0$. We show that, if $\tau (:= t - t_0) \to 0$, then p_t is a solution of the backward heat equation:

$$\partial_t p_t = -\triangle p_t, \quad p_{t=0} = p_0$$

where $\triangle$ denotes the Laplacian operator on M.

According to a fundamental formula for the transform of random variables,

$$p_0 = p_\tau \circ \Phi_\tau \cdot |\det \nabla \Phi_\tau|.$$

Clearly, if $t (= \tau) = 0$, then $p_{t=0} = p_0$.

By composing the asymptotic formula (4) and a Taylor expansion $p_\tau(x+\xi) = p_0(x) + \nabla p_0(x) \cdot \xi + \partial_t p_0(x) \tau + o(|\xi\tau|^2)$,

$$\begin{aligned} p_\tau \circ \Phi_\tau &= p_\tau \left[\mathrm{Id} + \tau \nabla \log p_0 + o(\tau^2) \right] \\ &= p_0 + \tau(\nabla p_0 \cdot \nabla \log p_0 + \partial_t p_0) + o(|\xi\tau|^2). \end{aligned}$$

On the other hand, by using a matrix calculus for the determinant of Hessian from Petersen and Pedersen [2012],

$$\begin{aligned} \det \nabla \Phi_\tau &= \det[I + \tau \nabla^2 \log p_0] \\ &= 1 + \tau \mathrm{Tr}\left[\nabla^2 \log p_0\right] + o(\tau^2) \\ &= 1 + \tau(p_0^{-1} \triangle p_0 - p_0^{-2} \nabla p_0 \cdot \nabla p_0) + o(\tau^2). \end{aligned}$$

Hence,

$$\begin{aligned} p_0 &= \left[p_0 + \tau(\nabla p_0 \cdot \nabla \log p_0 + \partial_t p_0)\right] \left[1 + \tau(p_0^{-1} \triangle p_0 - p_0^{-2} \nabla p_0 \cdot \nabla p_0)\right] + o(\tau^2) \\ &= p_0 + \tau(\nabla p_0 \cdot \nabla \log p_0 + \partial_t p_0 + \triangle p_0 - p_0^{-1} \nabla p_0 \cdot \nabla p_0) + o(\tau^2). \end{aligned}$$

Omitting higher order terms, it is reduced to the backward heat equation:

$$\partial_t p_0 = -\triangle p_0.$$

Finally, we show that the backward heat equation coincides with the final value problem for the heat equation. Let u_t be a solution of the final value problem

$$\partial_t u = \triangle u, \quad u_{t=\mathrm{T}} = p_0,$$

and $p_t^u := u_{\mathrm{T}-t}$. Then, $p_t^u = \Phi_{t\sharp} p_0$.

If $t = 0$, then $p_{t=0}^u = u_{\mathrm{T}-0} = p_0$. For $0 < t \leq \mathrm{T}$,

$$\begin{aligned}\partial_t p_t^u &= -\partial_s u_s\big|_{s=\mathrm{T}-t} \\ &= -\triangle u_s\big|_{s=\mathrm{T}-t} \\ &= -\triangle p_t^u.\end{aligned}$$

This concludes the claim.

C Proof of Proposition 5.3

By the assumption that p_0 is well-separated, we concentrate on the univariate Gaussian $p_0 = \mathcal{N}(0, \sigma_0^2)$ without loss of generality, because we can regard p_0 locally as a single component distribution, and diagonalize a single component Gaussian with a fixed matrix that is independent of time. According to (35), the standard deviation σ_ℓ monotonically shrinks as

$$\sigma_\ell = \left(1 + \frac{\tau}{\sigma_{\ell-1}^2}\right)^{-1} \sigma_{\ell-1}. \tag{30}$$

We show that (i) for every fixed τ, $\sigma_\ell \to 0$ with exponential decay, and (ii)

$$\lim_{\tau \to 0} \sigma_\ell^2 = \begin{cases} \sigma_0^2 - 2t & 0 \leq t \leq \sigma_0^2/2 \\ 0 & \text{otherwise} \end{cases}, \tag{31}$$

where $t := \ell\tau$.

Proof. The exponential decay is immediate by the monotonicity of σ_ℓ.

$$\sigma_\ell \leq \left(1 + \frac{\tau}{\sigma_0^2}\right)^{-1} \sigma_{\ell-1} \leq \left(1 + \frac{\tau}{\sigma_0^2}\right)^{-\ell} \sigma_0.$$

Write $a_\ell := \sigma_\ell^{-2}$. Then, (30) is reduced to

$$\begin{aligned} a_\ell &= (1 + a_{\ell-1}\tau)^2 a_{\ell-1} \\ &= a_{\ell-1} + 2a_{\ell-1}^2 \tau + o(\tau^2). \end{aligned}$$

Let $a(t)$ be a piecewise linear line that connects each a_ℓ at $t = \ell\tau$. In the limit $\tau \to 0$, we have

$$\frac{\mathrm{d}a}{\mathrm{d}t} = 2a^2, \quad a(0) = \sigma_0^{-2},$$

which is a separable equation and has the solution

$$a(t) = (\sigma_0^2 - 2t)^{-1}.$$

Observe that $1/a(t)$ becomes negative after $t = \sigma_0^2/2$. Hence, by the monotonicity with respect to ℓ and positivity of σ_ℓ, we can conclude (31).

D Denoising Autoencoder for Univariate Gaussians

We calculate the case for a univariate normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$.

D.1 Continuous Denoising Autoencoder

We show that

$$\varphi_t(x) = \sqrt{1 - \frac{2t}{\sigma_0^2}}(x - \mu_0) + \mu_0, \tag{32}$$

$$\varphi_{t\sharp}\mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}\left(\mu_0, \sigma_0^2 - 2t\right), \quad 0 \le t < \frac{\sigma_0^2}{2}. \tag{33}$$

Calculation of $\varphi_{t\sharp}$. Write the pushforward as $\mathcal{N}(\mu_t, \sigma_t^2)$. By using the heat kernel W_t, for some $\mathrm{T} > 0$,

$$\begin{aligned}\mathcal{N}(\mu_t, \sigma_t^2) &= \mathrm{W}_{\mathrm{T}-t} * \mathcal{N}(\mu_{\mathrm{T}}, \sigma_{\mathrm{T}}^2) \\ &= \mathcal{N}(\mu_{\mathrm{T}}, \sigma_{\mathrm{T}}^2 + 2(\mathrm{T} - t)).\end{aligned}$$

By eliminating T by the initial conditions, we have

$$\mathcal{N}(\mu_t, \sigma_t^2) = \mathcal{N}(\mu_0, \sigma_0^2 - 2t).$$

By the positivity of σ_t^2, we can determine the largest possible T as $\mathrm{T} = \sigma_0^2/2$.
Calculation of φ_t. Fix an arbitrary point x_0. Write $x_t := \varphi_t(x_0)$ and $\dot{x}_t := \partial_t \varphi_t(x_0)$. Recall that $\dot{\mu}_t \equiv 0$, because μ_t is a constant. According to (10),

$$\dot{x}_t = -\frac{x_t - \mu_t}{\sigma_t^2}.$$

By dividing both sides by x_t and integrating them,

$$\begin{aligned}\log\left|\frac{x_t - \mu_t}{x_0 - \mu_0}\right| &= -\int_0^t \frac{\mathrm{d}s}{\sigma_s^2} \\ &= \frac{1}{2}\int_0^t \frac{\mathrm{d}s}{s - \mathrm{T}} \\ &= \frac{1}{2}\log\left|\frac{\mathrm{T} - t}{\mathrm{T}}\right|.\end{aligned}$$

Hence, we have

$$x_t = \sqrt{1 - \frac{2t}{\sigma_0^2}}(x_0 - \mu_0) + \mu_0.$$

D.2 Discrete Denoising Autoencoder

We show that

$$\Phi_t(x) = \frac{\sigma_0^2}{\sigma_0^2 + t} x + \frac{t}{\sigma_0^2 + t} \mu_0, \tag{34}$$

$$\Phi_{t\sharp}\mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{(1 + t/\sigma_0^2)^2}\right). \tag{35}$$

Proof. The proof is immediate from (3). First,

$$\mathrm{W}_{t/2} * \mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}(\mu_0, \sigma_0^2 + t).$$

Hence,

$$\Phi_t(x) = x + t\nabla \log[\mathcal{N}(\mu_0, \sigma_0^2 + t)] = \frac{\sigma_0^2}{\sigma_0^2 + t} x + \frac{t}{\sigma_0^2 + t} \mu_0.$$

As Φ_t is affine, the pushforward is immediate.

E Denoising Autoencoder for Multivariate Gaussians

We calculate the case for a multivariate normal distribution $\mathcal{N}(\mu_0, \Sigma_0)$.

E.1 Continuous Denoising Autoencoder

We show that

$$\varphi_t(x) = \sqrt{I - 2t\Sigma_0^{-1}}(x - \mu_0) + \mu_0, \tag{36}$$

$$\varphi_{t\sharp}\mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}\left(\mu_0, \Sigma_0 - 2tI\right). \tag{37}$$

Proof. Recall that $\mathrm{W}_t * \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, \Sigma + 2tI)$. Thus, the pushforward $\mathcal{N}(\mu_t, \Sigma_t)$ is obtained as below in a similar manner to the univariate case.

$$\mathcal{N}(\mu_t, \Sigma_t) = \mathcal{N}\left(\mu_0, \Sigma_0 - 2tI\right).$$

Suppose that $\varphi_t(x)$ is an affine transform $A_t(x - \mu_0) + \mu_0$ as analogous to the univariate case. Recall that, if $X \sim \mathcal{N}(\mu, \Sigma)$, then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$. Hence, for our case $\Sigma_t = A_t \Sigma_0 A_t^\top$ and we can determine

$$A_t = \sqrt{\Sigma_t \Sigma_0^{-1}} = \sqrt{I - 2t\Sigma_0^{-1}}.$$

Finally, we check whether φ_t satisfies (10). As Σ_0 is symmetric, we can diagonalize $\Sigma_0 = U D_0 U^\top$ with an orthogonal matrix U and a diagonal matrix D_0. Observe that with the same U, we can simultaneously diagonalize Σ_t and A_t as

$$\Sigma_t = U D_t U^\top, \quad D_t := D_0 - 2tI$$
$$A_t = U D_t^{1/2} D_0^{-1/2} U^\top.$$

Without loss of generality, we can assume $U = I$ and therefore Σ_t and A_t are diagonal and $\mu_t \equiv 0$. Fix an index j and denote the j-th diagonal element of Σ_t and A_t by σ_t^2 and a_t, respectively. Then, our goal is reduced to showing $\partial_t[a_t x] = \nabla \log p_t(a_t x)$ for every fixed $x \in \mathbb{R}$.

By definition,

$$\begin{aligned}\sigma_t^2 &= \sigma_0^2 - 2t,\\ a_t &= \sigma_t \sigma_0^{-1} = \sqrt{1 - 2t\sigma_0^{-2}}.\end{aligned}$$

Thus, the LHS is

$$\partial_t[a_t x] = -\frac{1}{\sigma_0 \sqrt{\sigma_0^2 - 2t}} x = -\sigma_0^{-1}\sigma_t^{-1} x,$$

and the RHS is

$$\nabla \log p_t(a_t x) = -\frac{a_t x}{\sigma_t^2} = -\sigma_0^{-1}\sigma_t^{-1} x.$$

Hence, the LHS equals the RHS.

E.2 Discrete Denoising Autoencoder

We show that

$$\Phi_t(x) = (I + t\Sigma_0^{-1})^{-1} x + (I + t^{-1}\Sigma_0)^{-1}\mu_0, \tag{38}$$

$$\Phi_{t\sharp}\mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}\left(\mu_0, \Sigma_0 (I + t\Sigma_0^{-1})^{-2}\right). \tag{39}$$

Proof. Calc (3) directly as similar to the univariate case. First,

$$\mathrm{W}_{t/2} * \mathcal{N}(\mu_0, \Sigma_0) = \mathcal{N}(\mu_0, \Sigma_0 + tI).$$

Hence,

$$\begin{aligned}\Phi_t(x) &= x + t\nabla \log[\mathcal{N}(\mu_0, \Sigma_0 + tI)]\\ &= x + t\nabla\left[-\frac{1}{2}(x - \mu_0)^\top (\Sigma_0 + tI)^{-1}(x - \mu_0)\right]\\ &= (I + t\Sigma_0^{-1})^{-1} x + (I + t^{-1}\Sigma_0)^{-1}\mu_0.\end{aligned}$$

As Φ_t is affine, the pushforward is immediate.

F Denoising Autoencoder for Mixture of Gaussians

We calculate the case for the mixture of multivariate normal distributions $\sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k)$, with the assumption that it is *well-separated*.

F.1 Continuous Denoising Autoencoder

We show that

$$\varphi_t(x) \approx \sqrt{I - 2t\Sigma_k^{-1}}\,(x - \mu_k) + \mu_k, \quad x \in \Omega_k, \text{ if well-separated} \tag{40}$$

$$\varphi_{t\sharp} p_0 = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mu_k, \Sigma_k - 2tI\right), \tag{41}$$

where

$$\gamma_{kt}(x) := \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k - 2tI)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k - 2tI)}. \tag{42}$$

Proof. The pushforward is immediate by the linearity of the heat kernel. The dynamics (10) for our case is reduced to

$$\partial_t \varphi_t(x) = -\sum_{k=1}^{K} \gamma_{kt} \circ \varphi_t(x)(\Sigma_k - 2tI)^{-1}(\varphi_t(x) - \mu_k),$$

where γ_{kt} is given by (42).

By the assumption that p_0 is well-separated, we can take an open neighborhood Ω_k of μ_k and an open time interval I that contains t such that $\gamma_{kt} \circ \varphi_t(x) \equiv 1$ for every $(x, t) \in \Omega_k \times I$. In this restricted domain, the dynamics is reduced to a single component version:

$$\partial_t \varphi_t(x) = -(\Sigma_k - 2tI)^{-1}(\varphi_t(x) - \mu_k), \quad (x, t) \in \Omega_k \times I.$$

According to the previous results, we have exactly

$$\varphi_t(x) = \sqrt{I - 2t\Sigma_k^{-1}}(x - \mu_k) + \mu_k, \quad (x, t) \in \Omega_k \times I.$$

F.2 Discrete Denoising Autoencoder

We show that

$$\Phi_t(x) = \sum_{k=1}^{K} \gamma_{kt}(x) \left\{(I + t\Sigma_k^{-1})^{-1}x + (I + t^{-1}\Sigma_k)^{-1}\mu_k\right\}, \tag{43}$$

$$\Phi_{t\sharp} p_0 \approx \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mu_k, \Sigma_k (I + t\Sigma_k^{-1})^{-2}\right), \quad \text{if well-separated} \tag{44}$$

where

$$\gamma_{kt}(x) := \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k + tI)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k + tI)}. \tag{45}$$

Proof. Directly calc (3). By the linearity of the heat kernel,

$$\begin{aligned}\Phi_t &:= \mathrm{Id} + t\sum_{k=1}^{K} \frac{\pi_k \nabla \mathcal{N}(\mu_k, \Sigma_k + tI)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k + tI)}, \\ &= \mathrm{Id} + \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}(\mu_k, \Sigma_k + tI)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k + tI)} \cdot t\nabla \log \mathcal{N}(\mu_k, \Sigma_k + tI), \\ &= \mathrm{Id} + \sum_{k=1}^{K} \gamma_{kt}(\Phi_{kt} - \mathrm{Id}), \\ &= \sum_{k=1}^{K} \gamma_{kt}\Phi_{kt},\end{aligned}$$

where Φ_{kt} exactly coincides with the flow induced by an individual k-th component.

In order to estimate the pushforward, we introduce some auxiliary variables. Write $\pi(k) := \pi_k,\ \gamma(k \mid \cdot) := \gamma_{kt}(\cdot)$ and

$$\begin{aligned}p_t(\cdot \mid k) &:= \mathcal{N}(\mu_k, \Sigma_k + tI), \\ p_t &:= \sum_k \pi(k) p_t(\cdot \mid k).\end{aligned}$$

Let $\tau_k(\cdot \mid x)$ be a probability measure that satisfies

$$\int_M \tau_k(y \mid x) p_0(x \mid k) \mathrm{d}x = p_t(y \mid k).$$

Note that τ_k is not unique. Recall that by definition if $X \sim p_0(\cdot \mid k)$, then $Y = \Phi_{kt}(X) \sim p_t(\cdot \mid k)$. Hence, τ_k is a stochastic alternative to Φ_{kt}.

Consider a probability measure

$$\sigma(\cdot \mid x) := \sum_{k=1}^{K} \gamma(k \mid x) \tau_k(\cdot \mid x).$$

Clearly, this is a stochastic alternative to Φ_t. We show that

$$\int_M \sigma(y \mid x) p_0(x) \mathrm{d}x \approx p_t(y).$$

The LHS is reduced as

$$\begin{aligned}\int_M \sigma(y \mid x) p_0(x) \mathrm{d}x &= \int_M \sum_{k=1}^{K} \gamma(k \mid x) \tau_k(y \mid x) \sum_{\ell} \pi(\ell) p_0(x \mid \ell) \mathrm{d}x && (46) \\ &= \sum_{\ell} \pi(\ell) \sum_{k=1}^{K} \int_M \gamma(k \mid x) \tau_k(y \mid x) p_0(x \mid \ell) \mathrm{d}x. && (47)\end{aligned}$$

Suppose that $\gamma(k \mid x)$ is an indicator function of a domain Ω_k where $\int_{\Omega_k} p_0(\cdot \mid k) \approx 1$. Then,

$$\begin{aligned}(47) &\approx \sum_{\ell} \pi(\ell) \int_{\Omega_\ell} \tau_k(y \mid x) p_0(x \mid \ell) \mathrm{d}x \\ &\approx \sum_{\ell} \pi(\ell) p_t(y \mid \ell) = p_t(y).\end{aligned}$$

This concludes the claim.

References

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data generating distribution. *Journal of Machine Learning Research*, 15:3743–3773, 2014. URL `http://jmlr.org/papers/volume15/alain14a/alain14a.pdf`.

Fabio Anselmi, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations. *Theoretical Computer Science*, pages 1–10, 2015. doi: 10.1016/j.tcs.2015.06.048. URL `http://www.sciencedirect.com/science/article/pii/S0304397515005587`.

Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27*, pages 2654–2662, Montreal, BC, 2014. Curran Associates, Inc. URL `http://papers.nips.cc/paper/5484-do-deep-nets-really-need-to-be-deep.pdf`.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems 26*, pages 899–907, Lake Tahoe, 2013. Curran Associates, Inc. URL `http://papers.nips.cc/paper/5023-generalized-denoising-auto-encoders-as-generative-models`.

Yoshua Bengio, Éric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *Proceedings of The 31st International Conference on Machine Learning*, volume 32, pages 226–234, Beijing, China, 2014. JMLR W&CP. URL `http://jmlr.org/proceedings/papers/v32/bengio14.pdf`.

Jake Bouvrie, Lorenzo Rosasco, and Tomaso Poggio. On invariance in hierarchical models. In *Advances in Neural Information Processing Systems 22*, pages 162–170, Vancouver, BC, 2009. Curran Associates, Inc. URL `https://papers.nips.cc/paper/3732-on-invariance-in-hierarchical-models.pdf`.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: 10.1002/cpa.3160440402. URL `http://onlinelibrary.wiley.com/doi/10.1002/cpa.3160440402`.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. doi: 10.1109/TPAMI.2012.230. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6522407&tag=1.

Emmanuel Jean Candès. *Ridgelets: Theory and Applications*. PhD thesis, Stanford University, 1998.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, 2010. URL http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, revised edition, 2015. ISBN 9781482242386.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, Montreal, BC, 2014. Curran Associates, Inc. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527. URL http://www.mitpressjournals.org/doi/pdf/10.1162/neco.2006.18.7.1527.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations 2014*, pages 1–14, Banff, BC, 2014. URL http://arxiv.org/abs/1312.6114.

Diederik P. Kingma, Shakir Mohamed, Jimenez Danilo Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems 27*, pages 3581–3589, Montreal, BC, 2014. Curran Associates, Inc. URL http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf.

Noboru Murata. An integral representation of functions using three-layered betworks and their approximation bounds. *Neural Networks*, 9(6):947–956, aug 1996. ISSN 08936080. doi: 10.1016/0893-6080(96)00000-7. URL http://www.sciencedirect.com/science/article/pii/0893608096000007.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. doi: 10.1006/jfan.1999.3557. URL http://www.sciencedirect.com/science/article/pii/S0022123699935577.

Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A probabilistic theory of deep learning. Technical report, Rice University, 2015. URL https://arxiv.org/pdf/1504.00641v1.pdf.

Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook, Version: November 15, 2012. Technical report, 2012.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations 2016*, pages 1–15, San Juan, Puerto Rico, 2016. URL http://arxiv.org/pdf/1511.06434v2.pdf.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems 28*, pages 3546–3554, Montreal, BC, 2015. Curran Associates, Inc. URL http://papers.nips.cc/paper/5947-semi-supervised-learning-with-ladder-networks.pdf.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: explicit invariance during feature extraction. In *Proceedings of The 28th International Conference on Machine Learning*, volume ICML'11, pages 833–840, Belleview, WA, USA, 2011. ACM. URL http://www.icml-2011.org/papers/455_icmlpaper.pdf.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep Boltzmann machines. In *Proceedings of The 12th International Conference on Artificial Intelligence and Statics (AISTATS) 2009*, volume 5, pages 448–455, Clearwater beach, Florida USA, 2009. JMLR W&CP. URL http://www.jmlr.org/proceedings/papers/v5/salakhutdinov09a/salakhutdinov09a.pdf.

Jascha Sohl-Dickstein, Peter Battaglino, and Michael R. DeWeese. Minimum probability flow learning. In *Proceedings of The 28th International Conference on Machine Learning*, volume ICML'11, pages 905–912, Belleview, WA, USA, 2011. ACM. URL http://www.icml-2011.org/papers/480_icmlpaper.pdf.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265, Lille, France, 2015. JMLR W&CP. URL http://jmlr.org/proceedings/papers/v37/sohl-dickstein15.pdf.

Sho Sonoda and Noboru Murata. Sampling hidden parameters from oracle distribution. In *24th International Conference on Artificial Neural Networks (ICANN) 2014*, volume 8681, pages 539–546, Hamburg, Germany, 2014. Springer International Publishing. doi: 10.1007/978-3-319-11179-7_68. URL http://link.springer.com/chapter/10.1007%2F978-3-319-11179-7_68.

Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 2015. doi: 10.1016/j.acha.2015.12.005. URL http://www.sciencedirect.com/science/article/pii/S1063520315001748.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer-Verlag Berlin Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL http://www.springerlink.com/index/10.1007/978-3-540-71050-9.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142. URL http://www.mitpressjournals.org/doi/pdf/10.1162/NECO_a_00142.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of The 25th International Conference on Machine Learning*, volume ICML'08, pages 1096–1103, Helsinki, Finland, 2008. ACM. doi: 10.1145/1390156.1390294. URL http://portal.acm.org/citation.cfm?doid=1390156.1390294.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010. URL http://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf.