Effective Theory Building and Manifold Learning

David Peter Wallis Freeborn

November 24, 2024

Abstract

Manifold learning and effective model building are generally viewed as fundamentally different types of procedure. After all, in one we build a simplified model of the data, in the other, we construct a simplified model of the another model. Nonetheless, I argue that certain kinds of high-dimensional effective model building, and effective field theory construction in quantum field theory, can be viewed as special cases of manifold learning. I argue that this helps to shed light on all of these techniques. First, it suggests that the effective model building procedure depends upon a certain kind of algorithmic compressibility requirement. All three approaches assume that real-world systems exhibit certain redundancies, due to regularities. The use of these regularities to build simplified models is essential for scientific progress in many different domains.

1 Introduction

Manifold learning is a very widespread family of dimensional reduction techniques in machine learning, in which high-dimensional data is projected onto a lower-dimensional manifold, while preserving some salient properties of the original data (Belkin and Niyogi, 2001; Hinton and Roweis, 2002; Hinton and Salakhutdinov, 2006; McInnes et al., 2018; Roweis and Saul, 2000; Tenenbaum et al., 2000; van der Maaten and Hinton, 2008). This technique is based on the assumption that many high-dimensional datasets contain regularities that allow them to be conveniently compressed or summarized with a simpler model. Likewise, effective theory or model construction is a family of techniques in physics and the computational sciences, in which a high-dimensional theory or model is reduced to a lower-dimensional one. Effective theory building is commonly used in quantum field theory, where the mathematical problems have led to the construction of lower-dimensional effective field theories, and in many computational sciences, where there are many high-dimensional models, highly insensitive to the vast majority of parameter combinations (Burgess, 2020; Duncan, 2012; Machta et al., 2013; Raju et al., 2018; Transtrum et al., 2015).

Manifold learning and effective model building are generally viewed as fundamentally different types of procedure. After all, in one we build a simplified model of the data, in the other, we construct a simplified model of the another model (Monsalve-Bravo et al.,

2022; Quinn et al., 2022; Teoh et al., 2020). Indeed, they use the term **model** to mean two importantly different things.

- **Machine Learning Models** are functions that map from high-dimensional input data to lower-dimensional outputs, such as classifications or predictions.
- **Scientific/Computational Models** are mathematical representations of physical systems, typically mapping from theoretical parameters to observable predictions.

Nonetheless, I argue that certain kinds of high-dimensional effective model building, and effective field theory construction in quantum field theory, can be viewed as special cases of dimensional reduction techniques akin to manifold learning. I argue that this helps to shed light on some underlying principles shared by all of these techniques. First, it suggests that the effective model building procedure depends upon a certain kind of algorithmic compressibility requirement. All three approaches assume that real-world systems exhibit certain redundancies, due to regularities. The use of these regularities to build simplified models is essential for scientific progress in many different domains.

These topics have generated significant philosophical interest in recent years. There has been an ongoing debate over how effective theories and related methods can inform and refine scientific realism, particularly in the context of quantum field theory. Proponents of *effective realism* (Fraser, 2018,2,2; Miller, 2017; Wallace, 2006; Williams, 2019) argue that these methods can inform and refine a localized, theory-specific approach to realism by identifying the elements of quantum field theory (QFT) models that are empirically robust and likely to persist through scientific progress. However, this defense has been challenged by critics like Ruetsche (2018), who argue that while effective realism engages directly with successful aspects of current physics, it fails to fully mitigate skeptical challenges. Ruetsche suggests that these issues merely retreat to a different level rather than being resolved. Similarly, Rivat (2021) contends that effective theories rely on intrinsic empirical limitations and infinite idealizations that constrain their scope to offer reliable ontological commitments. He argues that these idealizations, while useful for making accurate predictions within certain domains, pose significant challenges for ensuring the stability and approximate truth of theoretical representations through future theory changes.

Likewise, philosophers have debated the the related topic, reduction and emergence in the context of renormalization group methods (see section 9). Batterman (2002,1) argues that phenomena such as critical behavior and phase transitions require explanations that transcend simple deductive reductions, emphasizing the importance of renormalization group theory in understanding how macroscopic properties emerge from microscopic interactions. He contends that the renormalization group theory reveals how different scales interact and influence each other, demonstrating that certain macroscopic behaviors cannot be fully reduced to microscopic laws. Similarly, Morrison (2012) highlights how renormalization group theory exemplifies the interplay between reduction and emergence in practice. Conversely, Butterfield (2014) proposes that reduction and emergence

are not mutually exclusive, arguing that these techniques provide a means to connect micro and macro levels, thereby reconciling reductionism with emergent properties.

In section **2**, I explain the dimensional reduction. I introduce manifold learning as a particular case of this in section **3**. In section **4**, I present the manifold hypothesis, and suggest one way to explicate it in a partly formal way. In section **5**, I introduce the sloppy models program. In section **6**, I argue that an effective model building technique, the manifold boundary approximation method can be viewed as akin to a special kind of manifold learning. In section **8**, I introduce effective field theories, and in section **9**, I argue that it can be related to both the sloppy models program and manifold learning. I conclude by drawing some overall analogies between these approaches.

2 Machine Learning and Dimensional Reduction

Imagine that a machine learning specialist wants to build an artificial intelligence tool for recognizing handwritten numerical digits. As input data, they train their tool on the MNIST (Modified National Institute of Standards and Technology) database, a large collection of handwritten digits commonly used for training various image processing systems. The training data contains 60,000 28×28 pixel images of handwritten digits ranging from 0 to 9 (LeCun et al., 1998, 2010). The aim is build a tool that can, in some sense, latch onto and generalize from key features of these handwritten digits, and which can then be applied to correctly interpret new handwritten images of digits, from outside of the training data.

In effect, the artificial intelligence tool serves as a *model* of the data. We can think of such a model as a function, f, from a real-valued vector of the 784 pixels in each image, to a vector of ten real-valued output classifications, giving some measure of how likely the model thinks it is that the image represents each possible digit 0-9,[1]

$$f : \mathbb{R}^{784} \to \mathbb{R}^{10}. \tag{1}$$

This general task, finding a function, mapping a real, N-dimensional data-vector to an $M < N$-dimensional output vector, is very common across machine learning. Indeed, almost any machine learning task can be represented as the task of finding a function of this form.[2] This is closely related to standard ways to think about model-building across the computational empirical sciences more generally.[3]

[1]For instance, in a Bayesian model, these real-valued output classifications could represent probabilities.

[2]For example, we can represent almost any predictive AI task (e.g. image classification, speech recognition, natural language processing tasks such as sentiment analysis and machine translation, recommender systems, medical diagnosis, financial forecasting, etc.) *or* generative AI task (e.g. text generation with a large language model or image generation with an adversarial network), as the task of finding a function of this form (Bishop, 2006; Goodfellow et al., 2016; Hastie et al., 2009; LeCun et al., 2015; Murphy, 2012; Vapnik, 1995).

[3]For instance, see Breiman (2001), or for related examples, see Gutenkunst et al. (2007), and see Sozou et al. (2017); Sullivan (2022); Williamson (2009) for some philosophical considerations.

Our machine learning specialist might not merely seek the most predictively accurate model; often they will also want the model to be simple. Simpler models usually make lower demands on computational resources for training, inference and application; the results may be more robust to small modifications; they may be easier to interpret or explain; and they may be less inclined to overfit the training data, allowing for better generalizability to new data. Furthermore, a variety of technical problems are known to arise when the dimensionality of the data is very high compared to the number of datapoints, resulting in the so-called *curse of dimensionality* (Bellman, 1957,6).[4]

Fortunately, the key features higher-dimensional real-world data can often be conveniently summarized by models with lower numbers of parameters. For instance, the salient variations in the MNIST handwritten digits might be summarizable by a much smaller number of factors or dimensions - rather than specifying each individual pixel, perhaps we can summarize them with a smaller number of identifiable curves, loops and lines. This task is at the heart of machine learning, algorithmic compression, and computational model-building more generally.

Thus, an obvious approach to simplify the model would be to first build a lower-dimensional model of the data. That is, instead of applying our model, $f : \mathbb{R}^N \to \mathbb{R}^M$, to the data directly, we could first reduce the dimensionality of the data with a model, $m : \mathbb{R}^N \to \mathbb{R}^K$, and then apply a simpler model, $g : \mathbb{R}^K \to \mathbb{R}^M$, with $N < K < M$. If the two processes give the same outputs, then we can think of f as the composition of g and m, as in figure **1**. However, in reality this is an unrealistic assumption: the two processes should give *almost* the same outputs, but some information will be lost when compressing the model. We call the high-dimensional space the **feature space** of the data, and the low-dimensional space the **latent space**. This process is now widespread in machine learning (see Fisher 1936; Izenman 1975; Pearson 1901 for some historical background to these techniques, and for contemporary examples, see Belkin and Niyogi 2001; Hinton and Roweis 2002; Hinton and Salakhutdinov 2006; Jolliffe and Cadima 2016; McInnes et al. 2018; Roweis and Saul 2000; Tenenbaum et al. 2000; van der Maaten and Hinton 2008).

The key is that our dimensional reduction model, m, must preserve certain salient local or global features of the data, even as it throws out some of the information contained in the original data. The salient features encoded in the data might vary, depending on the task at hand. They might include geometric properties (such as distances between data points, angles or local curvatures) or topological properties (including shape and connectivity features like clusters, holes, and loops). For instance, with our MNIST data, perhaps different ways of writing the same digit (like a closed '4' versus an open '  ') might form distinct subclusters within a larger cluster. Topological information could help in understanding the transition between different writing styles (for example, a curly '9' might continuously morph into a straight '  '). We define cost functions to measure how well certain salient properties of the data are preserved by the function m.

[4]Loosely speaking, the curse of dimensionality problems refer to the general observation that, as the dimensionality of the data grows, the volume grows so rapidly that finite data becomes sparsely distributed and increasingly orthogonal, making distance measures less able to extract useful information.

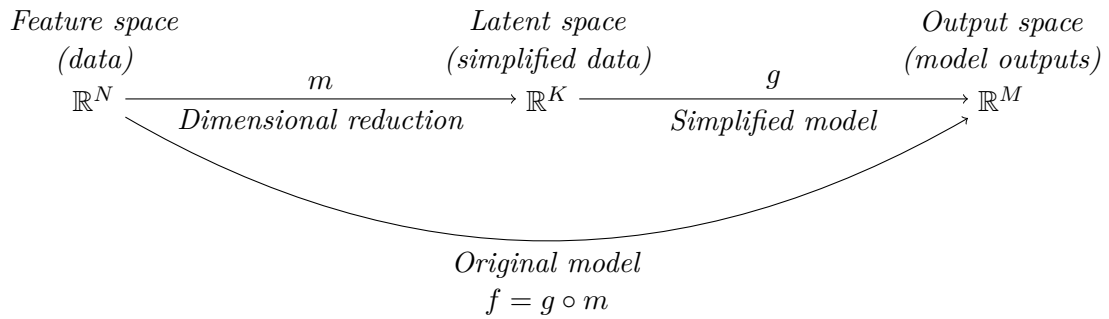


Figure 1: A category theoretic representation of the direct and simplified modeling approaches, assuming that they give the same outputs. Here, $\mathbb{R}^N$ is the feature space, $\mathbb{R}^K$ gives a latent space offering a simplified representation of the data, and $\mathbb{R}^M$ is the output space. The functions f, m, and g represent the original predictive model, the dimensionality-reducing model, and the simplified predictive model, respectively.

3 Manifold Learning

Roughly speaking, a manifold is a topological space that locally resembles flat Euclidean space.[5] Computer scientists have found that real-world data in $\mathbb{R}^N$ often lie close to a lower-dimensional manifold, $\mathcal{M}$, which can be embedded into $\mathbb{R}^K$, $K < N$. A suitable embedding function from $\mathbb{R}^N$ to $\mathbb{R}^K$ could provide a very convenient model of the data, one that preserves salient topological properties also being sensitive to nonlinear relationships between datapoints. In practice, manifolds seem to carry just the right amount of structure for this task (Belkin and Niyogi, 2001; Hinton and Roweis, 2002; Hinton and Salakhutdinov, 2006; McInnes et al., 2018; Roweis and Saul, 2000; Tenenbaum et al., 2000; van der Maaten and Hinton, 2008).

Roughly speaking, **manifold learning** is a family of dimensional reduction algorithms that progress according to the following scheme.

- We begin with the dataset with datapoints $x_i \in \mathbb{R}^N$. We posit that there exists a manifold $\mathcal{M}$ of dimension $K < N$, embedded in $\mathbb{R}^N$, such that the data points lie on or close to the manifold.
- The goal is to find an embedding function $m : \mathbb{R}^N \to \mathbb{R}^K$ that projects each high-dimensional datapoint in the feature space onto the K-dimensional latent space,

[5] More fully, an n-dimensional topological manifold is a topological space $\mathcal{M}$ which satisfies three conditions: First, it must be locally Euclidean, meaning that for every point p in $\mathcal{M}$, there exists an open neighborhood U around p that is homeomorphic to an open subset of Euclidean space $\mathbb{R}^n$, where n is a fixed integer representing the manifold's dimension. This ensures that sufficiently small neighborhoods in $\mathcal{M}$ locally resemble flat Euclidean space. Second, $\mathcal{M}$ must obey the Hausdorff condition, that for any two distinct points in $\mathcal{M}$, there exist disjoint open neighborhoods. This ensures that points can always be separated by open sets. Finally, $\mathcal{M}$ must be second-countable, meaning it possesses a countable basis for its topology. See Guillemin and Pollack (1974); Hirsch (1994) for further details.

$\mathbb{R}^K$.[6] It maps $\mathbb{R}^N$ onto $\mathbb{R}^K$ such that the images $m(x_i)$ preserve the intrinsic geometric and topological structure of the original data x_i on the manifold $\mathcal{M}$, within the constraints of the reduced dimensionality.

- We define a cost function,
$$C : \mathbb{R}^N \times \mathbb{R}^K \to \mathbb{R}, \tag{2}$$
which assigns a real number to each pair of points, one from the feature space and one from the latent space, designed to measure how well a map preserves salient geometric and topological features of the data (i.e. structural features of $\mathcal{M}$).

- We find an embedding, m, that minimizes the cost function.

- Finally, the reduced-dimension data points, y_i, are represented in the lower-dimensional latent space by their images under the embedding, $m(x_i) = y_i \in \mathbb{R}^K$.

When applying this procedure, it is essential to avoid **overfitting**, in which the model captures noise in the data, thereby failing to provide generalizable insights about the data. In the extreme case, without any procedures to avoid overfitting, we might represent all the data with a one-dimensional manifold, a curve passing through each datapoint. While this curve would perfectly 'fit' the data, it would fail to capture the simpler, underlying structures that we seek to learn

Therefore any manifold learning technique will generally require us to implement some techniques to prevent overfitting, often in the form of a smoothness constraint. There are three widely-used (non-exclusive) approaches to this.

- **Constraints on the manifold:** We explicitly restrict the class of allowable manifolds to those meeting certain smoothness criteria. I will discuss one example, the *reach constraint* in section **4**.[7]

- **Cost Function:** We favour smoother and simpler manifolds implicitly in the cost function. For example, in the *Locally Linear Embedding* algorithm (Roweis and Saul, 2000), we mitigate the effect of noise by approximating each point as a linear combination of its nearest neighbors. These nearest neighbours are likely to be part of the same smooth patch of the manifold. Overly contorted manifolds are often disfavored by this process.

- **Regularization:** We further modify the cost function to penalize insufficiently smooth solutions. Such a cost function might look like, $C_{\text{total}}(x_i, y_i) = C_{\text{base}}(x_i, y_i) + R(y_i)$, Where C_{base} measures how well the low-dimensional representation y_i preserves the structure of the original data x_i, and R is a regularization term that increases with the 'roughness' of the embedding (Hastie et al., 2009).

[6] The function m is an embedding if it is a smooth, injective, immersion, whose underlying continuous function is a homeomorphism onto its image (see Hirsch 1994, pages 21-29 for further details).

[7] These constraints are important in the theoretical studies of manifold learning. But they are not so widely used in practically useful algorithms, as directly enforcing such constraints can be computationally expensive (Belkin et al., 2006; Berenfeld et al., 2022; Fefferman et al., 2016).

Let us consider a very simple example of manifold learning (see Tenenbaum et al. 2000) and loosely show how to apply one possible local manifold-learning algorithm,[8] *Locally Linear Embedding* (Roweis and Saul, 2000). Suppose that our data is composed of points in a three-dimensional feature space, $\mathbb{R}^3$. Further suppose that the datapoints tend to lie close to a surface, described by the *swiss roll* parametric equations,

$$x^1 = t\cos(t) \tag{3}$$
$$x^2 = s \tag{4}$$
$$x^3 = t\sin(t), \tag{5}$$

where $x^{1,2,3}$ are some choice of the three coordinates, and t and s are parameters (see figure **2a**).[9]

Let us suppose that we want to reduce the dimensionality of this data to a latent space of just two dimensions, whilst trying to preserve the geometric features of the original global non-linear structure. If the datapoints lie near a two-dimensional manifold as we hope, then there should be a linear mapping from the coordinates of each neighbourhood to coordinates on the manifold which preserves this structure. So one approach could be to proceed as follows. First, we identify the k nearest neighbors for each point in the dataset, and some choice of integer, k, using on Euclidean distance in the $\mathbb{R}^3$ space. We assume that each data point and its neighbors lie to close to a *locally linear* patch of the manifold. Then, each point x_i can be reconstructed from a linear sum of the coordinates of its neighbor, $x_j s$. As such, we minimize the cost function, ε,

$$\varepsilon = |x_i - \sum_j W_{i,j} x_j|^2, \tag{6}$$

where the weights $W_{i,j}$ give the contribution of the jth data point to the ith reconstructed point.

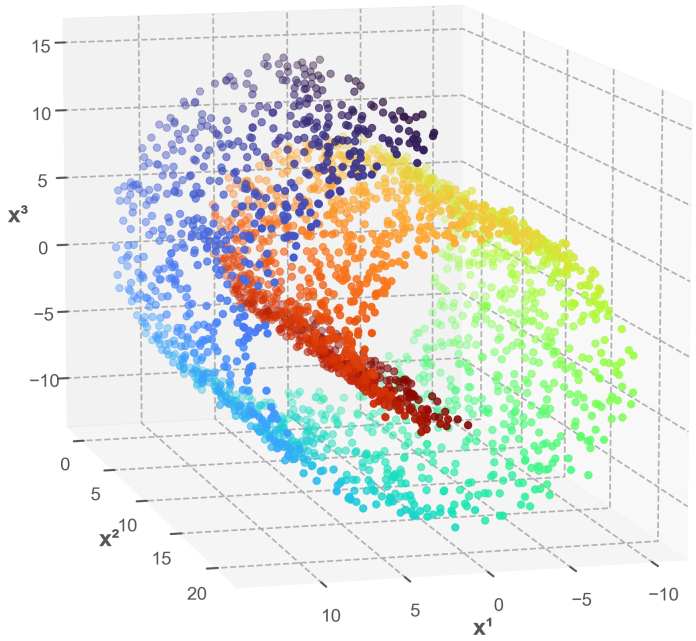
Finally, we find the corresponding points, y_i, in the latent space, $\mathbb{R}^2$, that best preserve these local weights. This is done by minimizing another cost function, ϕ,

$$\phi = |y_i - \sum_j W_{i,j} y_j|^2, \tag{7}$$

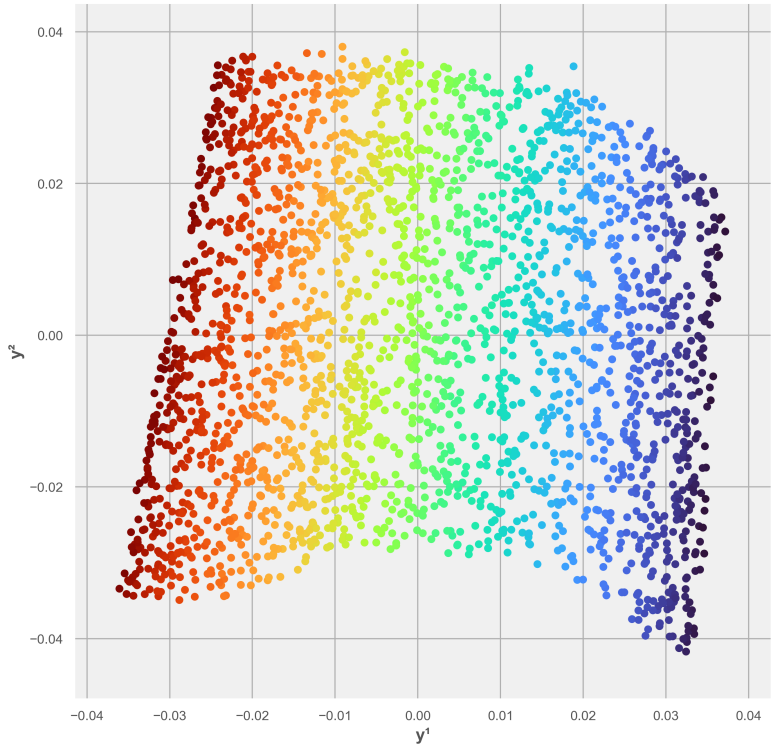
where y_i, y_j are the corresponding lower-dimensional embeddings of x_i and x_j respectively. The result is a lower-dimensional expression of the original data, preserving some of its original geometric features, albeit with some (hopefully small) loss of information. Figure **2b** shows the application of this algorithm to the data from figure **2a**: our swiss roll has been unfurled and flattened into a pancake.

[8] We can loosely distinguish two kinds of manifold learning algorithm, local and global methods (Cayton, 2005). For local methods, the cost function considers the placement of each point with respect to its neighbors, whereas for global methods tend to consider the relative placement of all points.

[9] The data was generated using the `scikit-learn` dataset, `make-swiss-roll` (Pedregosa et al., 2011).



(a) 2500 randomly generated datapoints in $\mathbb{R}^3$, lying close to the swiss-roll surface. The colours are for visualization only.



(b) The datapoints transformed under the LLE ($k = 20$) algorithm, represented in $\mathbb{R}^2$. Corresponding datapoints keep their colour from figure **2a**.

4 The Manifold Hypothesis

It is widely posited that all such manifold learning techniques share a common *fundamental assumption* (Cayton, 2005), often referred to as the *manifold hypothesis*. This assumption has rarely been stated rigorously. However, roughly speaking, it posits that high-dimensional real-world data can be sufficiently well-represented by data lying on a lower-dimensional latent manifold, embedded within the feature space (see Athanasopoulou et al. 2014; Bengio et al. 2013; Bordt et al. 2023; Brahma et al. 2016; Brown et al. 2022; Gorban and Tyukin 2018; Ivanov et al. 2021; Izenman 2012; Meilă and Zhang 2023; Narayanan and Mitter 2010; You and Ma 2011).[10]

It will be helpful to distinguish a *local* manifold hypothesis from a *global* manifold hypothesis. Given some dataset, the *local* manifold hypothesis states this dataset can be well-represented by data lying on a lower-dimensional latent manifold, embedded within the feature space. On the other hand, the *global* manifold hypothesis is the proposition that many real-world datasets can be effectively compressed by this kind of manifold learning; indeed that this is a prevalent feature of real-world datasets. One plausible and suitably general way to explicate the local manifold hypothesis could be as follows.[11]

Let $\mathcal{X} \subset \mathbb{R}^N$ be a high-dimensional feature space, with datapoints, $x_i \in \mathcal{X}$. Let $\mathcal{G}_{\mathcal{X}}(K, V, \tau)$ be the class of sub-manifolds in $\mathcal{X}$ with dimension, K, K-dimensional volume $\leq V$ and reach $\geq \tau$.[12] Then the manifold hypothesis is the assumption that, for some choice of $K < N, V, \tau$, there exists a manifold, $\mathcal{M} \in \mathcal{G}$, such that,

$$\mathcal{L}(\mathcal{M}, \{x_i\}) < \epsilon, \tag{8}$$

where $\mathcal{L}(\mathcal{M}, \{x_i\})$ is some measure of the average shortest distance (perhaps the mean-squared shortest distance) between the datapoints $\{x_i\}$ and the manifold $\mathcal{M}$, according

[10] Each individual manifold learning technique also makes a number of further assumptions. However this fundamental assumption, the manifold hypothesis is, by definition, shared by all manifold learning techniques.

[11] The main principles of this definition come from Fefferman et al. (2016). They define an algorithm to test the manifold hypothesis within a certain domain, for independent and identically distributed probabilistic data supported on a separable Hilbert Space. For our general purposes, it serves to loosen some of these requirements, whilst restricting ourselves to finite data on an N-dimensional space of real-numbers.

[12] We only want to consider manifolds above a certain reach and below a certain volume to avoid overfitting; after all, manifolds of sufficiently large volume or low reach could more easily capture every datapoint. Following Fefferman et al. (2016), the reach, τ, of a manifold is defined as the largest distance such that any point within the distance τ from the manifold has a unique closest point on the manifold. Sometimes loosely described as a measure of smoothness, one can more accurately think of it as a measure of local feature size, related to both local curvature and global *bottlenecks* (see Berenfeld et al. 2022 for a more complete explanation). The K-dimensional volume is given by the standard Lebesgue measure in $\mathbb{R}^K$. Recall (section **3**) that once we define a cost function for some manifold-learning algorithm, we might expect that such an overfitted manifold might nonetheless have a high cost, indicating that it does not properly capture the salient features of the data. In that sense, there is a risk of double-counting this requirement in this definition. One alternative would be to define a cost function from the outset, and require a manifold below a certain cost in the hypothesis. However, here I conservatively choose to stick to the approach used by Fefferman et al. (2016), defining the manifold hypothesis *prior* to specifying any cost function.

to some choice of distance (possibly, but not necessarily, the Euclidean distance in $\mathbb{R}^N$), and $\epsilon \in \mathbb{R}$ is some closeness threshold.

We can view the manifold hypothesis as an *data compressibility assumption*. The high-dimensional dataset contains redundancy. As such, the data can be well-represented with the use of a lower-dimensional model, without significant loss of information.

Expressed in this way, the local manifold hypothesis asserts that there exists a manifold $\mathcal{M}$ in $\mathcal{X}$ such that the average distance $\mathcal{L}(\mathcal{M}, P)$ is less than or equal to some specified threshold, chosen based on the desired level of proximity between the data distribution and the manifold. For different applications of the hypothesis, such as with different types of dataset or different manifold learning algorithms, we might choose to consider different manifold parameters, and different ways to measure the average distance and the threshold.

The global manifold hypothesis asserts that this applies to many real-world datasets. A general argument for the hypothesis has not been put forward; however, it has often been presented as a reason why machine learning is possible at all (see Cayton 2005; Fefferman et al. 2016; Olah 2014). After all, the higher the dimension of the data is, the harder machine learning tasks generally become. The global manifold hypothesis suggests that machine learning algorithms can potentially reduce the complexity of these tasks, by latching onto a smaller number of salient regularities in the data.[13] For instance, the task of interpreting handwritten digits in our MNIST dataset is far easier than one might naively fear from the high dimensionality of the data; handwritten versions of the same digits can be summarized by certain common, higher-level features (see Yao et al. 2017 for one example).

5 Sloppy Models in the Computational Empirical Sciences

Manifold learning has usually been applied to machine learning tasks, where we wish to build a simple model of the data for tasks like image recognition. Now let us turn to a framework in the computational empirical sciences,[14] known as sloppy modeling. Here, the task is to create a simple model of some target system, for the purpose of generating accurate predictions, and hopefully to better understand the system. However, we will see a strong analogy between this framework and machine learning.

In the computational empirical sciences, a model is a function from a real-valued vector of M' parameters to a real-valued vector of N' predictions $f' : \mathbb{R}^{M'} \to \mathbb{R}^{N'}$. Generally, we call $\mathbb{R}^{M'}$ the **parameter space**, and $\mathbb{R}^{N'}$ the **prediction space**. Often, the dimensions of the parameter space might represent properties of the system theoretically

[13] To take a more specific example, this principle is key to explaining the possibility of certain regularization techniques in deep learning, like dropout or weight decay, are effective (see Srivastava et al. 2014; Zou and Hastie 2005 for further details).

[14] I will use the term *computational empirical sciences* to refer to the wide array of scientific disciplines focused on using computational methods to build empirically-supported models of highly complex, high-dimensional target systems. This includes a wide array of fields, including, but not limited to, much of systems biology, chemistry, condensed matter physics, and many areas of engineering and the social sciences.

posited by our model, whereas the dimensions of the prediction space represent observable quantities that we measure. As such, observed datapoints also lie in the prediction space.

We can use measurements to estimate the model parameters. Measurements are represented by a set of real numbered vectors in the prediction space. We write a cost function to measure the distance between the model predictions and the empirical measurements, and tune the model parameters to minimize this cost. If our model is predictively accurate for a given choice of parameters, then the model's predictions should lie close to the measurements. Then we might then use such a model to generate further accurate predictions.

Consider this simple example.[15] Suppose we want to create a computational-scientific model of a pair of apparently identical pendulums, joined together with some string, and starting at rest but with one pendulum displaced (this is our target system). We can measure two things: the displacement of the first pendulum (x_1) or the displacement of the second pendulum (x_2), each indexed by different times, t (we can think of as t an independent regressor variable). So if we take measurements at ten different times, our measurement space will be $2 \times 10 = 20$ dimensional. Physicists often model such a system as a pair of weakly coupled, identical harmonic oscillators,

$$x_1(t) = d \cos\left(\frac{\omega_2 - \omega_1}{2} t\right) \cos\left(\frac{\omega_1 + \omega_2}{2} t\right), \tag{9}$$

$$x_2(t) = d \sin\left(\frac{\omega_2 - \omega_1}{2} t\right) \sin\left(\frac{\omega_1 + \omega_2}{2} t\right). \tag{10}$$

where we have three parameters: ω_1 and ω_2 are normal frequencies of the system, representing the frequency of the pendulums oscillating with the same amplitude in phase and out of phase, and d is the initial displacement of one pendulum. Note that these natural frequency parameters are theoretical posits of our model, which we do not directly measure. Rather, we infer their values using our model and observations of the displacements.

A model is described as **sloppy** if its predictions are highly insensitive to most parameter combinations (which we call the sloppy parameter combinations), but are highly sensitive to a small number of parameter combinations (which we call the stiff parameter combinations) (see Quinn et al. 2022; Transtrum et al. 2015 for scientific overviews and see Freeborn 2024 for a philosophical analysis). This allows for significant alterations in the values of sloppy parameter combinations, potentially by factors in the thousands or tens of thousands, with minimal impact on the model's predictive output. Thus a model with M' parameters might operate with a considerably lower *effective* dimensionality in practice. Following Freeborn (2024), we call a physical target system sloppy if it can be well-represented by a sloppy model, i.e. if we can produce an effective model that is a good description of the system. We could operationalize as the requirement that the datapoints lie close to an effective model manifold.

[15] See Pain 2005, pages 79-87 for a fuller treatment.

We can measure the sensitivity of the specific parameter combinations to the observed data using the Fisher Information Matrix (FIM). This gives the expected curvature of the log-likelihood function of the observed data in relation to the model parameters. The eigenvectors of the FIM are termed local or "renormalized" eigenparameters.[16]

Many real-world systems seem to depend on huge numbers of parameters. However, it becomes increasingly hard to build good models with large numbers of parameters. Just as in machine learning, high-dimensionality can be a major problem in computational scientific modeling. The utility of a sloppy model lies in its ability to effectively capture the salient features of a dataset while demonstrating a robust tolerance to variations in many of its parameters. As such, sloppy systems are suitable targets for **effective models**, in which some or all sloppy parameter combinations can be ignored. Fortunately, scientists have found that systems across a very wide variety of domains are sloppy, ranging from systems biology to quantum mechanics to particle accelerators (Gutenkunst et al., 2007). Proponents of the sloppy models framework argue that the ubiquity of sloppy systems can help to explain the success of science (Freeborn, 2024; Quinn et al., 2022; Transtrum et al., 2015).

Hopefully, by expressing things in this way, the analogy with machine learning is already clear. Note that the scientific model's prediction space, $\mathbb{R}^{N'}$, corresponds to the machine learning model's feature space, $\mathbb{R}^{N}$. In each case, the dimensions of the space correspond to the real-world observable quantities in the target system; a measurement of each of these quantities corresponds to a datapoints in that space. The scientific model's parameter space, $\mathbb{R}^{M'}$, corresponds to the machine learning model's latent space, $\mathbb{R}^{M}$. The dimensions of this space correspond to higher-level theoretical quantities (parameters or latent variables) posited by our model of the target system. However, observe that the function f' in the computational empirical sciences takes an opposite direction to the function f in machine learning. The former takes us from our model parameters to the observable predictions, whereas the latter takes us from observable data in the feature space to tune the predicted model parameters. It will often be helpful to assume that such functions are invertible in both the machine learning and computational sciences contexts.

The effective model m' is a function from a simplified, lower-dimensional space of $K' < M'$ effective (or "renormalized") parameters, $\mathbb{R}^{K'}$ to the prediction space, $\mathbb{R}^{N'}$. As we will see in there next section, we can also propose a *manifold boundary approximation function*, g', to take us from the high-dimensional, to the low-dimensional prediction space. If the effective model and original sloppy model make the same predictions, then

[16]The Fisher Information Matrix gives the expectation of the second-order partial derivatives of the log-likelihood function of the observed data with respect to the model parameters. Viewing this Fisher Information "matrix" as a metric (a type (0,2) tensor), it is given by,

$$g_{\mu\nu}(y') = \mathbb{E}\left[\frac{\partial \log p(x'|y')}{\partial y'^{\mu}} \frac{\partial \log p(x'|y')}{\partial y'^{\nu}}\right] \tag{11}$$

where $\mu, \nu = 1, 2, \ldots, M'$, y' is the M'-dimensional parameter vector, x is the N-dimensional predictions vector, $p(x'|y')$ gives the likelihood of observing the predictions x' given the parameters y' in the model, and $\mathbb{E}[\cdot]$ gives the expectation with respect to the distribution of the observed measurements.

the relation between these functions is shown in figure **3**. Once again, in reality this is an unrealistic assumption: the two models should give *almost* the same outputs, but some information will be lost in the effective model.

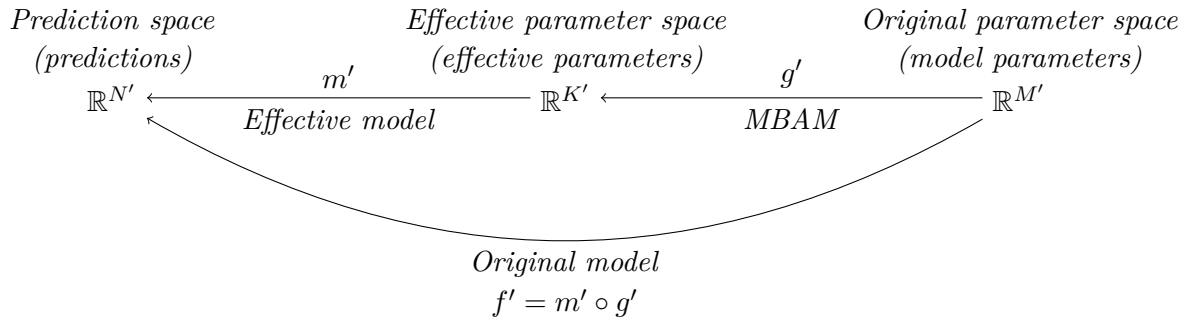


Figure 3: A category theoretic representation of the original and effective modeling approaches, assuming that they give the same outputs. Here, $\mathbb{R}^{N'}$ is the prediction space, $\mathbb{R}^{K'}$ is the effective parameter space, and $\mathbb{R}^{M'}$ is the original sloppy parameter space. The functions f', g', and m' represent the original sloppy model, the manifold boundary approximation method (MBAM) function, and the effective model function, respectively.

6 Manifold Boundary Approximation in the Computational Sciences

We can derive these effective models by using an information-geometric approach (see Transtrum et al. 2011,1), in which we endow the model with a little more structure. Each vector of predictions defines a point in the prediction space, $\mathbb{R}^{N'}$, and each vector of model parameters y' generates one such point under the function f'. As such (assuming $M' > N'$), we can reinterpret the model as an M'-dimensional sub-manifold $\mathcal{R}'$, embedded in the prediction space, $\mathbb{R}^{N'}$. This embedded sub-manifold is defined by the points,

$$\mathcal{R}' = \{f(y') \in \mathbb{R}^{N'} : \text{ for all the parameter combinations } y' \in \mathbb{R}^{M'}\}. \tag{12}$$

Here y' gives the manifold coordinates: as such, varying the parameters y' of the model moves along the manifold surface, leading to a different point (vector of predictions) in the feature space in which it is embedded. The collection of all these points (for all possible parameter values) forms the model manifold surface.[17]

The FIM can serve as a Riemannian metric on the model manifold, measuring parameter space distances (in units of standard deviations of the parameter, given their

[17]To interpret the model as an embedding, we must make some further assumptions about f'. It must be smooth, injective, an immersion, and its underlying continuous function must be a homeomorphism onto its image (see Hirsch 1994, pages 21-29 for further details).

probability distributions under the model). Such distances operationalize the distinguishability between model predictions from different parameter choices.

We can explore how the model predictions change as we vary corresponding parameter combinations by tracing geodesics along the model manifold. If we move far enough along a geodesic, we may eventually reach a point where further movement would take us to boundaries. Beyond these boundaries, the model's predictions become non-physical, undefined, or irrelevant.[18] For instance, such boundaries can arise when certain parameter combinations are not physically meaningful, or lead to singularities or mathematically undefined behavior.

Tthe existence of these boundaries on the model manifold represents a general principle of model reduction. This concept suggests that simpler models often arise at the extremes of parameter values, an idea implicit in many areas of physics and elsewhere in the computational sciences. The modern framing of this as 'manifold boundaries' provides a rigorous mathematical foundation for this intuition.

Therefore, geodesic lengths give an indicator of sloppiness: long geodesics correspond to stiff parameter combinations, whilst short geodesics correspond to sloppy parameter combinations, in which the values can be varied over many orders of magnitude without significantly altering the model predictions in $\mathbb{R}^{N'}$. The shape of a sloppy manifold is described as a *hyperribbon*, with many short dimensions and only a few longer dimensions.

There are various ways to identify and utilize these boundaries for model reduction. One method designed for this purpose is the **manifold boundary approximation method** (MBAM). This uses an information-geometric approach to systematically find lower-dimensional effective models for sloppy systems (Transtrum and Qiu, 2014). This allows us to propose a manifold boundary approximation function, $g' : \mathbb{R}^{M'} \to \mathbb{R}^{K'}$, taking vectors of the original parameters to corresponding vectors of the effective parameters. The MBAM algorithm proceeds as follows to find a particular boundary of the model manifold.

- We begin with the dataset with datapoints $x'_i \in \mathbb{R}^{N'}$, and the embedded model manifold, $\mathcal{R}'$. We posit that the system is sloppy.
- The goal is to find an embedding function $g' : \mathbb{R}^{M'} \to \mathbb{R}^{K'}$ that projects each high-dimensional parameter vector in the parameter space onto the K'-dimensional manifold, $\mathcal{M}$. We seek an effective model, $m' : \mathbb{R}^{K'} \to \mathbb{R}^{N'}$, such that for each parameter vector within some chosen domain, y'_i, $f'(y'_i) \approx m'(g'(y'_i))$, i.e., the sloppy model and the effective model should make approximately the same predictions within this domain.
- We find an embedding, g', as follows:

[18] A K-dimensional manifold with boundary $\mathcal{M}$ is a topological space where every point p in $\mathcal{M}$ has a neighborhood homeomorphic to an open set in the Euclidean half-space $\mathbb{R}^K_+ = \{(x_1, \ldots, x_K) \in \mathbb{R}^K : x_K \geq 0\}$. Points in $\mathcal{M}$ that have neighborhoods homeomorphic to an open set in $\mathbb{R}^n$ (the entire Euclidean space) are called *interior points*. Points in $\mathcal{M}$ that have neighborhoods homeomorphic to an open set in $\mathbb{R}^K_+$ but not in $\mathbb{R}^K$ are called *boundary points*.

– **Model Fitting:** Initially, fit the model to data to find a best-fit point in the parameter space using an appropriate cost function,

$$C : \mathbb{R}^{N'} \times \mathbb{R}^{M'} \to \mathbb{R}, \tag{13}$$

which assigns a real number to each pair of points, one from the prediction space and one from the parameter space.[19]

– **Eigenvalue Analysis:** Compute the FIM at this point and perform an eigenvalue analysis. Identify the direction associated with the smallest eigenvalue, which corresponds to the sloppiest parameter combination.

– **Geodesic Tracing:** Trace a geodesic in the parameter space along this sloppiest direction. This path leads to a boundary of the model manifold where the insensitive parameter becomes redundant.

– **Model Reduction:** At the manifold boundary, effectively remove or fix the redundant parameter, thereby reducing the model's dimensionality.

– **Effective Model Fitting:** The effective model parameters are fit to the data using an analogous cost function,

$$C : \mathbb{R}^{N'} \times \mathbb{R}^{K'} \to \mathbb{R}, \tag{14}$$

which assigns a real number to each pair of points, one from the prediction space and one from the effective parameter space, designed to measure how well a map preserves salient geometric and topological features of the data.

– **Iteration:** Repeat the process as needed to simplify the model further, focusing each time on the next sloppiest direction.

- Finally, the effective parameter vectors, y'_i, are represented in the lower-dimensional effective parameter space by their images under the embedding, $g'(y'_i) = y'_i \in \mathbb{R}^{K'}$.

Recall the problem with overfitting in manifold learning discussed in section **3**. The MBAM procedure is generally resistant to such overfitting. Unlike manifold learning techniques, which might find arbitrary lower-dimensional representations to fit data, MBAM is constrained to move along existing model structures. The boundaries it finds correspond to limiting cases of the original model, preserving its core structure rather than introducing new, potentially overfitting parameters. This process is guided by the model's intrinsic geometry, not by fitting to specific data points. Thus MBAM does not introduce new complexity to match particular observations.

To illustrate a simplified, toy version of this procedure, consider our coupled pendulums model once more. This model is not sloppy *in general*; however, it can illustrate some principles effective model building because some of its parameter combinations can become sloppy in certain domains, such as when the times t are very small. Here, we

[19]The task of fitting the model parameters can be interpreted as projecting the data onto the model manifold (Quinn et al., 2022).

shall focus on just the displacement of the first pendulum, x_1. First, let us rewrite the model in terms of some new parameter combinations, $\omega_{\mathrm{h}} = \omega_2 + \omega_1$ and $\omega_{\mathrm{l}} = \omega_2 - \omega_1$. Now the displacements x_1 and x_2 are characterised by a higher frequency (ω_{h}) modulated sinusoidal oscillation, varying within a lower frequency (ω_{l}) sinusoidal envelope (figure **4a**). At small enough times, t, we might find that the model predictions are highly insensitive to the value of ω_l; only the high frequency variations seem to matter: in this domain, ω_l has become sloppy. We might find that the geodesics reveal a model boundary at $\omega_l \to 0$. At this boundary, corresponding to small times and small couplings, we can eliminate the dependence on ω_l, and rewrite our model,

$$x_1(t) = d\cos\left(\omega_{\mathrm{h}} t\right), \tag{15}$$

which we can think of this as a effective model of the system adequate at a particular domain, with just two parameters, d and ω_{h} (figure **4b**).

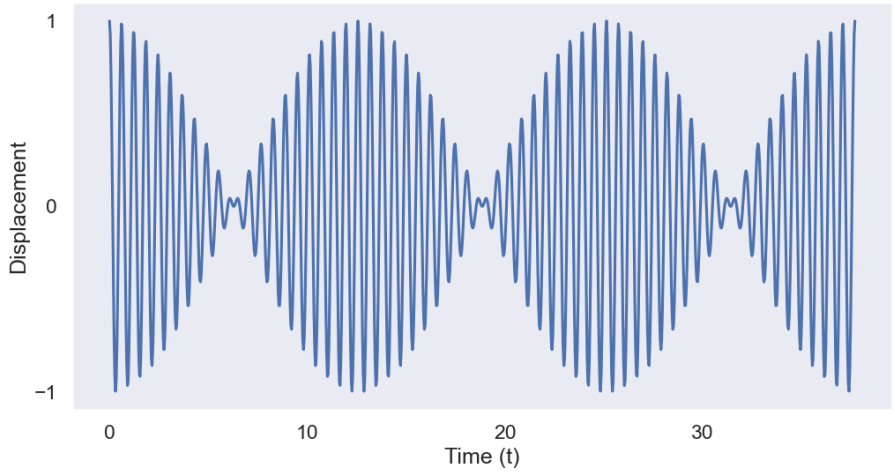
7 Manifold Boundary Approximation and Manifold Learning

Prima facie, MBAM appears to be a fundamentally different procedure to Manifold Learning, as described in section **3**. There is some truth to this. After all, manifold learning provides a way to build a simplified model of the data, taking us from datapoints, x_i, in the data space, $\mathbb{R}^N$, to prediction vectors, $y_i = m(x_i)$, in a reduced latent space, $\mathbb{R}^K$. On the other hand, the Manifold Boundary Approximation Method takes us from parameter vectors y'_i in the parameter space, $\mathbb{R}^{M'}$ to effective parameter vectors, $y'_i = g'(y'_i)$, in the effective parameter space, $\mathbb{R}^{K'}$. Recalling our analogy between machine learning and the computational empirical sciences, the data space corresponds to the prediction space, the latent space corresponds to the parameter space, and the reduced latent space corresponds to the effective parameter space. As the diagrams in figures **1** and **3** demonstrate, these are not the same procedure.[20]

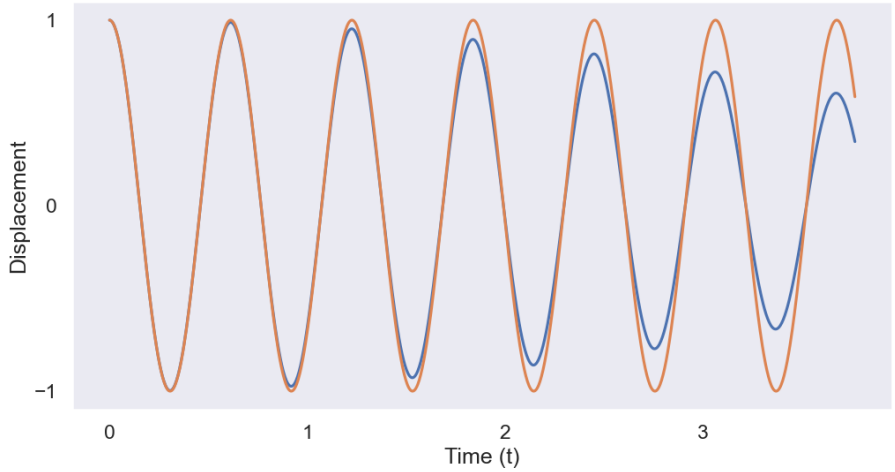
However, recall that before starting the MBAM procedure, we reinterpreted the model $f'(y'_i) = x'_i$ as an M'-dimensional manifold, $\mathcal{R}'$, with y' giving the manifold coordinates, embedded in the prediction space, $\mathbb{R}^{M'}$. We begin the MBAM procedure by finding the parameter vector on this manifold, best tuned to the datapoints, x'_i, and then follow a geodesic to reach the manifold boundary. Finally, the sloppy parameter combination is eliminated, and the effective model is re-tuned to the data. In effect, this procedure identifies a K'-dimensional sub-manifold of the dataspace, $\mathbb{R}^{N'}$, that is tuned to the data, just as with the case of manifold learning.

To be clear, the procedures are not identical here. Under manifold learning, we find the specific sub-manifold (for a given K) that is best tuned to the data according to the cost function. By contrast, the manifold boundary found by following a sloppy geodesic is in no sense guaranteed to be the K'-dimensional sub-manifold best tuned to the data

[20]Indeed, they have generally been viewed as two essentially distinct kinds of procedure (Monsalve-Bravo et al., 2022; Quinn et al., 2022; Teoh et al., 2020).



(a) Predicted displacement of the first pendulum according to the original model, for some choice of parameter values.



(b) Predicted displacement of the first pendulum with the same parameter values, at small times, according to the original model (blue) and effective model (orange). Notice that the two models almost agree at sufficiently small times.

according to any given cost function. Rather than simply opting to find the best sub-manifold, in the MBAM procedure, we first seek to eliminate the sloppiest parameter combinations. These are precisely the parameter combinations that are least sensitive to the data, and in general likely to be hardest, and least relevant, to tune. The purpose of eliminating such parameter combinations is precisely to make the task of tuning the effective model easier. Whilst MBAM does not entail finding the sub-manifold most tuned to the data, it does find the sub-manifold whose most *stiff* parameter combinations are best tuned to the data.

Here it is worth considering the two procedures epistemically. When performing manifold learning, we begin with knowledge of the data, $x_i \in \mathbb{R}^N$. We do not assume knowledge of the prior model f or of the latent variables we wish to measure that make up $\mathbb{R}^M$. The manifold learning procedure uses the data to give us the function m that can best represent the datapoints (and in effect the function g), and the reduced latent space, $\mathbb{R}^N$. Now consider MBAM. We begin with analogous knowledge, of the data, $x'_i \in \mathbb{R}^{N'}$, but also require prior knowledge of the model function, f' and the model parameters that make up $\mathbb{R}^{M'}$, . The manifold learning procedure uses these to give us the function g' (and in effect the function m' that allows us to best represent the datapoints using the stiff parameters) and the effective parameter space, $\mathbb{R}^{K'}$.

Thus, both procedures construct the same kind of object, a sub-manifold of the feature space or prediction space, $\mathbb{R}^N$ or $\mathbb{R}^{N'}$, tuned appropriately to the datapoints $x_i \in \mathbb{R}^N$ or $x'_i \in \mathbb{R}^{N'}$. However, there are two key epistemic differences. First, when performing MBAM, we start with greater knowledge: we must already have some prior model of the data, f', unlike in manifold learning. The epistemic role of f' is precisely to help us identify the stiff parameter combinations in constructing $\mathbb{R}^{K'}$. Further note that identifying these is the *only* epistemic role of the model f': the effective parameter tuning takes place with respect to the data, in precise analogy to manifold learning. Second, when we perform MBAM, we restrict ourselves to best representing the data only using these stiff parameter combinations.

To stress this point, let us consider the MNIST dataset example once more. Suppose that our machine learning specialist has finally found a model, f, taking the real-valued vector of 784 pixles in each time to a vector of ten-real-valued output classifications. However, they find the model overly complex and difficult to tune, due the many apparently insensitive parameters. Perhaps the task would be easier if they could find a reduced dimensionality model, which doesn't depend on all these parameters. They use the data to identify the insensitive parameter combinations, and then build a effective model of the data, $g : \mathbb{R}^K \to \mathbb{R}^N$. Observe how this process relates closely to the process of building a dimensional reduction model, described in section **3**, albeit utilising already-existing model.

Confusion is natural given the different ways functions have been defined in machine learning and the computational empirical sciences. However, whilst the methods differ, it perhaps makes sense to think of the MBAM procedure as akin to a special kind of manifold learning procedure, in which a prior model, f' is used to identify the stiff parameter values, and then the reduced latent manifold (or effective parameter space)

is fit to the data only using the stiff parameter combinations.

This motivates a reconsideration of *sloppiness* itself. Following Transtrum et al. (2010); Transtrum and Qiu (2014), we can describe a model $f' : \mathbb{R}^{M'} \to \mathbb{R}^{N'}$ as sloppy if $\mathcal{R}'$ (the corresponding M'-dimensional sub-manifold of $\mathbb{R}'$) has few stiff geodesics, and many sloppy geodesics. Each such sloppy geodesic must have at least one manifold boundary. The MBAM algorithm provides one way to identify and utilize these boundaries, allowing us to eliminate sloppy parameter combinations and build an effective model of reduced-dimensionality. If we can build an effective model with $M' - K'$ fewer parameters, then this model will correspond to a K'-dimensional sub-manifold of $\mathbb{R}^{N'}$.

So, one plausible explication of sloppiness would be with the following criterion. Let $\mathcal{X}' \subset \mathbb{R}^{N'}$ be a high-dimensional prediction space, with datapoints, $\{x'_i\} \in \mathcal{X}'$. Require that $\mathcal{Y}' \subset \mathbb{R}^{M'}$ is a high-dimensional parameter space, with an embedding $f' : R^{M'} \to \mathbb{R}^{N'}$ defining a sub-manifold in the prediction space, $\mathcal{R}' \subset \mathbb{R}^{N'}$. Let ${\mathcal{G}'}_{\mathcal{X}'}(K', V, \tau)$ be the class of *effective model* sub-manifolds in $\mathcal{R}'$ with dimension, K', volume $\leq V$ and reach $\geq \tau$. Then the sloppiness criterion demands that, for some choice of $K' < N', V, \tau$, there exists an effective manifold, $\mathcal{M}' \in \mathcal{G}'$, such that,

$$\mathcal{L}(\mathcal{M}', \{x'_i\}) < \epsilon, \tag{16}$$

where $\mathcal{L}((\mathcal{M}, \{x_i\})$ is some measure of the average shortest distance between the datapoints $\{x'_i\}$ and the manifold $\mathcal{M}$, according to some choice of distance, and $\epsilon \in \mathbb{R}$ is some closeness threshold.

Observe that this sloppiness criterion is essentially the manifold hypothesis from section **4**, alongside an addition requirement: the sub-manifold that describes the data must be an effective model manifold, produced from a sloppy model. This additional requirement restricts the scope of the models under consideration, but in a reasonable way: finding an effective model precisely involves eliminating parameter combinations which are least sensitive to the data.

Another way to think of this is that manifold learning might be a *more* theory-independent way of building a dimensional reduction model of the data. However, as noted, any particular manifold learning algorithm will add additional assumptions. Finding an effective model is *more* theory-dependent: we begin with an assumption that a prior model f' can offer a good description of the system, but has some superfluous (sloppy) parameter combinations which can be eliminated, reducing the dimensionality of the model. However, above this assumption, the additional assumptions of effective-model building are quite weak, namely that the improved model of the system involves precisely removing those irrelevant parameter combinations.

We could imagine MBAM's goal as reducing the number of parameters, while retaining as much relevant information from the original model as possible, ensuring that the essential features and behaviors of the system are preserved. We could imagine a "cost function" in terms of trace of the FIM, to measure the fidelity of the reduced model to the original model's manifold. Against such a cost, then MBAM would behave like a *greedy algorithm*[21], making locally optimal choices at each step without guaranteeing a

[21] A greedy-algorithm iteratively makes the locally optimal choice at each step, typically without

globally optimal solution. As such, MBAM does not explore the whole space of solutions.

We could consider other algorithms for achieving this goal. At the most constrained end of the spectrum, MBAM represents a greedy approach, in which it finds a particular K'-dimensional boundary through iterative geodesic tracing. While computationally efficient, this does not guarantee finding the globally optimal reduced model. By contrast, manifold learning seeks to find the best K-dimensional representation of the data without constraints on the form of that representation, and requires explicit smoothness assumptions to avoid overfitting. Less constrained alternatives to MBAM might be algorithms that find the optimal K'-dimensional edge or boundary of the model manifold, or even choosing a K'-dimensional sub-manifold of the original model manifold, without the requirement that it be an edge or boundary. Such algorithms could still inherit structural constraints on the model, and therefore might not require a further procedure to prevent overfitting. Such algorithms could plausibly have useful properties relevant to both scientific model-reduction and machine learning.

Just like the manifold hypothesis, we can view sloppiness as a *compressibility assumption*. The high-dimensional dataset and the model contain redundancy. As such, the data can be well-represented with the use of a lower-dimensional effective model, without significant loss of information.

8 Effective Field Theories in Physics

The sloppy models program is closely tied to the modern effective field theories (EFTs) program in quantum field theory and particle physics.[22] It is therefore worth considering to what extent the effective field theories program in physics can also be understood through the use of manifold learning methods.

EFTs are built on the principle that quantum field theoretic phenomena decouple at different energy or distance scales. This means that to describe physics at a certain scale, it is only necessary to consider relevant degrees of freedom and interactions at that scale. Furthermore, given our ignorance about physics at certain scales, it is often necessary to do so. As with sloppy models, EFTs simplify the description of complex systems by identifying the most relevant degrees of freedom and parameters at a given scale.[23]

Many highly successful theories are effective field theories, including Fermi's Theory of Weak Interaction, Chiral Perturbation Theory, Heavy Quark Effective Theory, and Ginzburg-Landau Theory. Indeed, on the modern approach pioneered by Wilson and Kogut (1974), we should expect *every* quantum field theory to be an effective field theory, including the Standard Model of particle physics.

backtracking. Such an algorithm may not find a global optimum.

[22]Indeed, at least some cases of effective field theory building can be understood as instances of effective model building in sense meant in sections **5** and **6** (See Machta et al. 2013; Raju et al. 2018; Transtrum et al. 2015 for further details and Freeborn 2024 for a philosophical discussion.

[23]For further details see Binney et al. 1992, Weinberg 1996 and Duncan 2012. For a philosophical overview, see Butterfield 2014 and Butterfield and Bouatta 2014 for further details).

We can think of a quantum field theory as a scientific model along essentially the same lines as the computational models in section **5**. We can define a space of predictions, $\mathbb{R}^{N'}$, consisting of measurable quantities, perhaps including physical quantities such as scattering amplitudes, cross-sections or decay rates; and a space of parameters, $\mathbb{R}^{M'}$ perhaps including the field variables and their so-called *bare* coupling constants, such as masses and charges.[24] We can interpret the theory as a function, $f' : \mathbb{R}^{M'} \to \mathbb{R}^{N'}$, mapping from a selection of parameters to a set of predictions. The theory is usually written in terms of a Lagrangian, $\mathcal{L}(\phi_i, \lambda_j)$, where ϕ_i are posited field variables, and λ_j are the field parameters. We can derive the equations of motion, and eventually the predictions about observables, through a rather involved process, starting with the Lagrangian.

Unfortunately, calculations using non-trivial interacting quantum field theories are typically found to lead to problematic mathematical divergences. To tackle these divergences, physicists modify these theories through a family of correction techniques known as **regularization**. For example, a simple way to do this is to impose a momentum cutoff scale, at a much higher energy than the interactions we wish to study. These corrections can render the theory finite, but usually lack a principled physical motivation.

The solution, **renormalization**, involves adjusting the bare parameters of the theory to remove dependence on any regularization scale. The resulting *renormalized theory* has new *renormalized parameters* that shift with the energy or distance scale, at which we describe the theory. We call the shifting of these renormalized *running* coupling constants the **renormalization group flow** (RG flow) through the parameter space. We derive differential equations to describe how the renormalized couplings must vary with the scale, if we impose the requirement that the physical observables in the prediction space must remain the same.

Crucially, the parameter trajectories under RG flow will remain *within* the parameter space of the theory. These parameter trajectories often lead to fixed points or surfaces,[25] at which the value of the renormalized coupling constants cease to change with the scale. Many different theories, those with the same fields and symmetries may flow towards the same fixed surface.[26] As such, the fixed surfaces are said to define *universality classes* of theories that share the same behavior at some scale.

Hence, near these fixed regions, the theory can be thought of as exhibiting a kind of self-similar behavior across scales, an approximate scale-invariance. In effect, the renormalizable part of the theory can be approximately decoupled from the physics energy scales. A linear approximation of the RG flow equations near stable fixed points reveals a small number of unstable directions, in contrast to the majority that are stable. Unstable directions correspond to parameters for which small changes can result in large

[24] We might think of both the field variables *and* field parameters as parameters of the theory. The theory posits both the *kinds* of fields (for example, scalar, vector, spinor etc), expressed through the field variables, and their bare coupling constants.

[25] There are generally two kinds of fixed points – Gaussian (or free-field) fixed points, where interactions vanish, and non-Gaussian fixed points, where the interactions reach a non-zero constant value. These fixed points are also crucial in understanding critical phenomena in phase transitions.

[26] For a proof in the case of one simple scalar theory, see Polchinski, 1984.

changes to the theory's predictions. These correspond to **relevant** and **marginal** parameters. Stable directions correspond to parameters for which small changes only lead to small changes to the theory's predictions. These correspond to **irrelevant** parameters.[27] Therefore, the predictions of the theory become dominated by a smaller number of *relevant* and *marginal* parameters.

In consequence, we can construct a lower-dimensional **effective field theory** by eliminating the irrelevant renormalized parameters. Such effective field theories provide a good model of the system at certain energy scales, usually at low energy, but break down at other scales, often high energy scales. As such, we can construct predictively successful, effective low-dimensional, low-energy theories even whilst remaining ignorant of the physics at higher energy scales.

Let us consider a simple example, an imaginary, simple scalar quantum field theory with two interacting fields, ϕ_L and ϕ_H with different masses, m_L and m_H respectively, with $m_L \ll m_H$. Suppose that we want to find a low-energy effective theory, relevant to sclaes $\Lambda \ll m_H$. We summarize the original theory with the Lagrangian,

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi_L)^2 - \frac{1}{2}m_L^2\phi_L^2 + \frac{1}{2}(\partial_\mu \phi_H)^2 - \frac{1}{2}m_H^2\phi_L^2 - \lambda\phi_L^2\phi_H^2, \tag{17}$$

where ∂_μ represents the partial derivative with respect to spacetime coordinates and λ is the coupling constant for the interaction between the two fields, ϕ_L and ϕ_H. At low energies, we find that the parameters associated with high-mass particles become irrelevant, and the lower mass fields become effectively decoupled from them. The Lagrangian relates to the field configurations of the theory by means of the partition function,

$$Z = \int \mathcal{D}\phi_L \mathcal{D}\phi_H \, e^{i\int d^4x \, \mathcal{L}(\phi_L, \phi_H)}. \tag{18}$$

We can eliminate the higher mass degrees of freedom by integrating over them and defining a new effective Lagrangian, $\mathcal{L}_{\text{eff}}$, as follows.[28]

$$Z = \int \mathcal{D}\phi_L \, e^{i\int d^4x \, \mathcal{L}_{\text{eff}}(\phi_L)}. \tag{19}$$

Unfortunately, in general such an effective Lagrangian may be characterized by an infinite series of terms,

$$\mathcal{L}_{\text{eff}} = \frac{1}{2}(\partial_\mu \phi_L)^2 - \frac{1}{2}m_L^2\phi_L^2 + \sum_{n=1}^{\infty} \frac{c_n}{m_2^n}\mathcal{O}_n(\phi_1), \tag{20}$$

[27] Roughly speaking, relevant parameters are those that whose effect on the theory's predictions increases close to a stable fixed point. Irrelevant parameters are those that whose effect on the theory's predictions decreases close to a stable fixed point. Other parameters are described as marginal (see Goldenfeld (1992, pages 245-246) for further details).

[28] There are two essential steps to this process. First, the Applequist-Carrazone decoupling theorem shows that the degrees of freedom associated with high-mass particles are suppressed at low energies (see Appelquist and Carrazone 1975 for further details) Second, the degrees of freedom can be separated and integrated out, to form an effective field theory (see Wilson (1983); Wilson and Kogut (1974) for further details).

where c_n are coefficients that depend on the details of the full theory, including the coupling constant, λ. Fortunately, in this case, power-counting considerations[29] and renormalization group arguments can show that these terms become irrelevant in the low energy limit we are interested in, resulting in a well-defined low-energy effective field theory, with Lagrangian $\mathcal{L}_{\text{eff-low-energy}}$,

$$\mathcal{L}_{\text{eff}} = \frac{1}{2}(\partial_\mu \phi_L)^2 - \frac{1}{2} m_L^2 \phi_L^2. \tag{21}$$

In this case, the low-energy theory looks exactly as we might have expected, with only the free fields for the low mass field, and no coupling to the high mass field. At this energy scale ($\Lambda \ll m_H$) and the high mass field is *frozen out*. The direct production of heavy particles associated with the high mass field energetically unfeasible, and their indirect effects, such as contributions to quantum corrections, are also negligible. Our effective model of the system describes free, low-mass fields.

Clearly, we can understand this overall procedure of finding a lower-dimensional effective field theory as a form of dimension reduction of our original theory's parameter space. Observe how similar this approach was to finding an effective model of the coupled oscillators in sections **5** and **6**.

9 Renormalization and Compressibility

For some simple models, renormalization group flow towards fixed regions can be recovered as a special case of MBAM. Indeed, just as the manifold hypothesis has been proposed as a reason why machine learning is possible, and sloppy modeling has been credited as an explanation for the success of science, many cite the renormalization group procedure as an explanation for successful theory building in high energy physics (Fraser, 2018,2,2; Miller, 2017; Wallace, 2006; Weinberg, 1996; Williams, 2019). This raises the question: can we say something stronger about the relationship between effective theory building in renormalized theories and the other manifold learning procedures we have discussed so far?

Here, it will serve to step back and consider the renormalization group in a more general setting than just quantum field theory. Recall from section **8**, that the renormalization group involved transforming parameters within the parameter space under a change of energy or some other scale, without changing the theory's predictions in the prediction space. The renormalization group transformation can be understood as a coarse-graining procedure, in which the short-distance degrees of freedom of the model are integrated out, effectively viewing the system with less and less precision. In order to keep the model predictions in agreement, these scale transformations require us to transform between points in the parameter space *and*, correspondingly, to rescale the prediction space.

[29] The operators built from fields (in this case, ϕ_L) can be organized according to the size of their contribution in a systematic expansion. See Burgess (2020, pages 51-81) for further details.

Like the MBAM procedure, RG flow does not generally face overfitting problems of the type we discussed in section **3**. Unlike data-driven dimensional reduction techniques, which might find arbitrary lower-dimensional representations to fit data, RG flow is constrained by the structure of the underlying theory. The transformations involved in RG flow are guided by the symmetries of the system and the requirement of scale invariance at fixed points. These transformations require that the effective models bear a self-similarity in form to the original model (except for a possible reduction of parameters). Ideally, each step in the RG flow corresponds to a physically meaningful transformation of the theory, such as integrating out short-distance degrees of freedom to obtain an effective theory. Once again, the process is guided by the theory's structure, not by fitting to specific data points.

Consider the following simple example (originating with Kadanoff, 1966). In the one-dimensional zero-field Ising model, we consider an infinite chain of coupled spins. A single parameter, J gives the coupling between neighbouring spins; the spins s_i at sites indexed by i, take values of ± 1 and give the predictions of the model. The model is often summarized using the Hamiltonian,

$$H = -J \sum_i s_i s_{i+1} \tag{22}$$

We can coarse-grain or "renormalize" the model by averaging out short-range details and focusing on long-range behavior. One way to do this is through the Kadanoff *block spin transformation*, in which we group spins into blocks of spins and then sum over the spins within each block to define new, effective spin variables. If we use blocks of 2 spins, the averaged spin of the block could be,

$$S_{\text{new}} = \text{sign}(s_1 + s_2). \tag{23}$$

After this process, we effectively reduce the number of degrees of freedom in our prediction space: we are now sensitive to only half as many spins as before. Perhaps now, we keep only the predictions for the spins at even sites in our prediction space. This might inspire us to write a new effective model of the system, in which the effective Hamiltonian, H', will only involve these spins,

$$H' = -J' \sum_i s_{2i} s_{2i+2}, \tag{24}$$

and in which the new effective coupling, J', serves as an effective coupling between what were previously blocks of spins. Requiring that the new model spin predictions correspond to the predictions of the original model, and assuming a probability distribution over the spins, we can derive a functional form of J' in terms of J, and derive renormalization group flow equations to understand how the coupling changes with scale. Successive iterations of the transformation lead towards fixed points, in which the effective couplings do not change further. Observe then that the transformation of the

parameter space, whilst keeping the effective predictions in correspondence with the previous predictions necessitates a loss of sensitivity to some of the degrees of freedom in the prediction space. In a sense, the effective model does not map to as many predictions.

More generally, assuming the computational modeling framework from sections **5** and **6** it will help to define the renormalization group flow as any transformation on the embedded model manifold, $\mathcal{R}'$, or within the parameter space $\mathbb{R}^{M'}$ under some change of scale, which coarse grains the model's predictions. Renormalization group transformations therefore require that the effective models bear a self-similarity in form to the original model (except for a possible reduction of parameters). As such, any coordinate-invariant geometric and topological features of the model manifold will remain fixed. This raises a question: if the model manifold does not change, and renormalization simply constitutes a flow along its surface, how is it that the renormalized model will lose degrees of freedom in the prediction space, $\mathbb{R}^{N'}$?

The key is to realize that the manifold's metric, given by the FIM, representing the distinguishability between model predictions from different parameter choices decreases. Raju et al. (2018) show that we can quantify the loss of information from discarding these degrees of freedom discarded through coarse-graining by finding how the metric tensor changes under a coarse-graining application. Let us specify a continuous coarse-graining procedure, where as the smallest length scale, $l = l_0 \exp(b)$, changes, the parameters change according to $\frac{dy'^{\mu}}{db} = \beta_{\mu}$, where y'^{μ} are the parameters, and β are the beta functions, which define the flow so as to preserve the predictions. Then the change in the metric under this flow is given by a modified Lie derivative, $\mathcal{L}_{\beta}$,

$$\mathcal{L}_{\beta} g_{\mu\nu} = \beta^{\alpha} \partial_{\alpha} g_{\mu\nu} + g_{\alpha\mu} \partial_n u \beta^{\alpha} + g_{\alpha\nu} \partial_m u \beta^{\alpha} - L \partial_L g_{\mu\nu}, \tag{25}$$

where the first term represents the directional change of the metric, and the second and third terms represent the change in the parameter space distances, as the parameters shift. The fourth term arises from the coarse-graining of the model, assuming that the size of the observed system length L shinks, according to $\frac{dL}{db} = -L$.[30] Raju et al. (2018) find that the metric decreases along the irrelevant directions, whilst it is preserved along the relevant and marginal directions.[31]

This motivates a careful consideration of the renormalization group flow procedure. This procedure is reminiscent to MBAM, but with two key differences. First, during renormalization group flow, we do not necessarily travel along the geodesics corresponding to the sloppiest parameter combinations. More crucially, we have added the additional step of course graining the predictions of the theory. This yielded a decrease in the FIM metric in certain directions, effectively increasing the sloppiness of the theory. Indeed, the parameters of such systems have been found to become increasingly sloppy as they approach fixed points under renormalization group flow (Machta et al., 2013; Raju et al., 2018). This increase in sloppiness corresponds to a loss of information as we

[30] As L is not a parameter, we must supplement the usual Lie derivative with this fourth term.

[31] This observation has been corroborated by computer simulations (Machta et al., 2013). Note that, in a closely related model, (Strandkvist et al., 2020) demonstrate that the changes in the metric due to the model deformation exactly corresponds to the changes in the metric induced by parameter flow.

perform renormalization group flow, precisely corresponding to the loss in sensitivty to certain degrees of freedom in the prediction space.[32]

Therefore, one way to understand the renormalization group flow procedure as we have defined it, is as a particular kind of transformation within the parameter space that also coarse grains the prediction space, thereby modifying the model to increasing sloppiness. It is precisely those irrelevant and sloppy parameters that we remove when building an effective theory. Insofar as effective theory building in physics involves increasing model sloppiness and then creating an effective model in the sense of section **6**, then this too could be interpreted as akin to a special kind of manifold learning. If we understand the renormalization group procedure as increasing sloppiness, then the task of effective theory building is straightforwardly analogous to the construction of an effective model of an (at least somewhat) sloppy system. This kind of effective theory construction seems to depend upon a criterion of theoretical compressibility.

10 Conclusions

Manifold learning, the sloppy models program, and effective field theories operate in three different scientific domains. Nonetheless, there are strong analogies between the three fields. Of course, all three share a basic principle in common: they seek to reduce the dimensionality of some data, model, or theory. However, more fundamentally, MBAM shares a close analogy with manifold learning. The two techniques have generally been seen as fundamentally different: manifold learning begins with a high-dimensional dataset and seeks to produce a low-dimensional model of it, whereas MBAM also seeks to build a lower-dimensional effective model of an already-existing high-dimensional model. However, both ultimately produce a low-dimensional model of the data: the difference is that MBAM uses greater prior knowledge to do so, in particular using a prior model to help identify the sloppy parameter combinations to remove. As such, I have argued that MBAM can be viewed as akin to a special kind of manifold learning.

Likewise, effective theory building in physics bears a close relationship to manifold learning. The renormalization group procedure can be understood as being in some ways analogous to MBAM, transforming the parameters of the theory. However, by simultaneously applying a coarse-graining to the predictions of the theory, it more drastically transforms the model, increasing the model's sloppiness. As such, effective theory building in physics could also be understood as akin to a special kind of manifold learning.

The manifold hypothesis underpins large areas of research in machine learning. If the global manifold hypothesis is right, then it may contribute to an explanation of why machines are capable of learning from complex data. Likewise, the sloppiness of real-world systems, or the existence of fixed points under renormalization group flow may contribute to explanations of why we can build scientific models of highly complex real-world systems. I have argued that all these assumptions share a basic common form: though the systems in question are superficially complex, they contain redundancy in

[32] For an alternative, but potentially related sense in which the renormalization group flow corresponds to a loss of information, see Zomolodchikov (1986).

the form of regularities. As such, the systems can be *compressed*. We can construct lower-dimensional effective models of the system by latching onto these regularities.

These technical connections between manifold learning, sloppy models, and effective field theories may have implications for several key philosophical debates, especially regarding the renormalization group. Recall that Batterman argues that some renormalization group phenomena are not reducible to lower-level theories. He contends that renormalization group techniques reveal how macroscopic properties emerge from microscopic interactions in ways that resist traditional forms of reduction due to the need for idealizations, such as the thermodynamic limit or infinite size assumptions. Viewing these methods through the lens of dimensional reduction techniques described in sections **2**, **3**, and **6** may help to shed some light. Each technique involves idealizations such as the successive elimination of sloppy parameter combinations at manifold boundaries in the MBAM procedure, or the reduction in dimensionality from the feature space to the latent space in manifold learning. These idealizations are cases of the kinds of infinite idealization discussed in Batterman (2002,0,1): we let certain parameters or parameter combinations shrink indefinitely small, to capture essential behaviors of physical systems.

Nonetheless, such dimensional reduction techniques seem to take the explicit form of approximate reductions, akin to that suggested by Butterfield (2014) in the context of the renormalization group.[33] In a traditional reduction scheme (Nagel, 1961; Schaffner, 1967) two key conditions must be met: there must be bridge laws that systematically correlate the theoretical terms of both theories, and the laws of the reduced theory must be logically derivable from the reducing theory's laws combined with the bridge laws and any necessary auxiliary assumptions. The dimensional reduction schemes discussed in this paper seem to take precisely this form. As we have seen, the dimensional reduction and simplified model can be understood as functions mapping from the feature space to the latent space, and from the latent space to the output space. Likewise, MBAM and the effective model can be understood as functions mapping from the original parameter space to the effective parameter space, and from the effective parameter space to the prediction space. The relations in figures **1** and **3** seem to suggest a form of dependence where higher and lower level models retain an explicit functional relationship.

Of course, in practice there *is* a loss of information here: the process is not generally reversible, and so figures **1** and **3** are themselves idealizations. After all, as we have seen, the key feature of these techniques is precisely such a reduction in the complexity of the models, and thus a loss of information. Renormalization group techniques explicitly reduce the information from the underlying theory, as we have seen in section **9**. Likewise, MBAM techniques reduce complexity by eliminating sloppy parameter combinations as we approach the manifold boundaries, retaining core predictive power in a lower-dimensional model without needing exact, fully derivable connections to microscopic details. These seem naturally understood as cases of approximate reduction, which can

[33]Recall that Butterfield argues that, in the context of the renormalization group, reduction and emergence are not mutually exclusive, suggesting that even phenomena that appear emergent can often be reconciled with a form of reduction when idealizations are understood as approximations rather than ontological separations.

maintain essential dependencies across scales without requiring strict derivability. Dimensional reduction techniques also exhibit structural stability, capturing scale-invariant features in complex data. For example, in manifold learning lower-dimensional embeddings retain the essential topology and geometry of data. However, such idealizations are most naturally viewed as pragmatic simplification rather than barriers to reduction, which are made explicit in the functional forms relating the parameter spaces of the various models. As such, manifold learning seems highly amenable to interpretation as a family of approximate Nagelian reductions (see Dizadji-Bahmani et al. 2010 for a defense of this model of reduction). They seem to provide a family of test cases in which infinite idealizations do not present a natural barrier to reductive techniques. It would be an interesting line for future research to demonstrate this directly in particular examples, and to reconcile such a reductive approach with the idealizations involved.

Furthermore, the techniques discussed in this paper suggest that some of the philosophical debates around the renormalization group may apply more widely across other sciences. Recall that Fraser (2018,2,2); Miller (2017); Wallace (2006); Williams (2019) have argued that the success of effective field theories supports a form of selective scientific realism. My contention that effective theory construction can be understood as a special case of nonlinear dimensional reduction techniques suggests that such arguments can be applied more widely. It suggests that the success of effective theories is not unique to physics, but reflects a more general feature of successful scientific modeling: the ability to identify and preserve essential features while eliminating irrelevant degrees of freedom.

However, the machine learning perspective presented in this paper does not, in itself provide a defense against these kind of skeptical challenges provided by Ruetsche (2018) and Rivat (2021). For example, suppose that we wish to adopt a scientific realist perspective on some effective parameter combinations in our effective theory. We have a good reason for doing so: the theory leads to accurate predictions **and** we have reason to believe that this effective theory will remain a good effective model, even if our knowledge of the original model parameters changes. However, if our underlying theory changed more drastically, for example requiring entirely different parameters, there is no guarantee that this effective model will remain effective. In essence, effective theories remain vulnerable to *unconceived alternative* theories, that lie outside of the model space under consideration (see Freeborn, 2024; Stanford, 2010).

If something akin to the global manifold hypothesis can be broadly defended, this might plausibly make room for a wider family of effective realist defenses of scientific theories, applying well beyond the scope of physics. After all, if the global manifold hypothesis holds, then many real-world datasets can be effectively compressed by dimensional reduction methods. As such, we might expect effective theory building techniques to be broadly successful precisely because many real-world target systems are amenable to them. Unconceived alternative theories might still replace our current best theories, but there would perhaps be less reason to expect them. One might defend it with an inference to the best explanation: the manifold hypothesis is the best explanation for the remarkable success of manifold learning across a wide variety of domains. However,

unfortunately, the global manifold hypothesis lacks a compelling theoretical motivation, with the main arguments being empirical Brahma et al. (2016); Fefferman et al. (2016); Gorban and Tyukin (2018).

Nonetheless, these techniques suggest one promising path for the selective realist in particular fields: to show that the salient epistemic features relevant to effective realism can also apply to a wider family of algorithmic reduction approaches, coupled with a suitable defense of a relevant local manifold hypothesis. Putting such an argument on a solid and rigorous footing would require substantial further work, but this suggests a potentially fruitful avenue for further research. There is not room to rigorously develop and defend such an argument here; however, Freeborn (2024) suggests one such possibility for extending effective realistic arguments beyond their traditional domain of quantum field theory.

References

Appelquist, T. and Carrazone, J. (1975). Infrared singularities and massive fields. *Physical Review D*, 11:2856.

Athanasopoulou, G., Iosif, E., and Potamianos, A. (2014). Low-dimensional manifold distributional semantic models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 731–740.

Batterman, R. W. (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press.

Batterman, R. W. (2005). Critical phenomena and breaking drops: Infinite idealizations in physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 36(2):225–244.

Batterman, R. W. (2011). Emergence, singularities, and symmetry breaking. *Foundations of Physics*, 41(6):1031–1050.

Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 585–591, Cambridge, MA, USA. MIT Press.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Berenfeld, C., Harvey, J., Hoffmann, M., et al. (2022). Estimating the reach of a manifold via its convexity defect function. *Discrete Computational Geometry*, 67:403–438.

Binney, J., Dowrick, N., Fisher, A., and Newman, M. (1992). *The Theory of Critical*

Phenomena: An Introduction to the Renormalization Group. Oxford Science Publ. Clarendon Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bordt, S., Upadhyay, U., Akata, Z., and von Luxburg, U. (2023). The manifold hypothesis for gradient-based explanations. In *IEEE/CVF 2023 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3697–3702.

Brahma, P. P., Wu, D., and She, Y. (2016). Why deep learning works: A manifold disentanglement perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 27(10):1997–2008. Epub 2015 Dec 7.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. (2022). The union of manifolds hypothesis. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*.

Burgess, C. P. (2020). *Introduction to Effective Field Theory: Thinking Effectively about Hierarchies of Scale*. Cambridge University Press.

Butterfield, J. (2014). Reduction, Emergence and Renormalization. *J. Philos.*, 111:5–49.

Butterfield, J. and Bouatta, N. (2014). Renormalization for philosophers. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 104.

Cayton, L. (2005). Algorithms for manifold learning. Technical Report 12(1–17), University of California at San Diego.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2010). Who's afraid of nagelian reduction? *Erkenntnis*, 73(3):393–412.

Duncan, A. (2012). *The Conceptual Framework of Quantum Field Theory*. Oxford University Press, Oxford, UK.

Fefferman, C., Mitter, S., and Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Fraser, J. D. (2018). Renormalization and the formulation of scientific realism. *Philosophy of Science*, 85(5):1164–1175.

Fraser, J. D. (2020a). The real problem with perturbative quantum field theory. *British Journal for the Philosophy of Science*, 71(2):391–413.

Fraser, J. D. (2020b). Towards a realist view of quantum field theory. In French, S. and Saatsi, J., editors, *Scientific Realism and the Quantum*, pages 276–292. Oxford University Press, Oxford, UK.

Freeborn, D. (2024). Sloppy models, renormalization group realism, and the success of science. *Erkenntnis*, pages 1–29.

Goldenfeld, N. (1992). *Lectures On Phase Transitions And The Renormalization Group*. CRC Press, 1st edition.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press.

Gorban, A. N. and Tyukin, I. Y. (2018). Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philosophical Transactions A: Math-*

ematical, Physical and Engineering Sciences, 376(2118):20170237.

Guillemin, V. and Pollack, A. (1974). *Differential Topology*. Prentice-Hall.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, 3(10):1–8.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics.

Hinton, G. and Roweis, S. (2002). Stochastic neighbor embedding. In *Neural Information Processing Systems*.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313:504–507.

Hirsch, M. W. (1994). *Differential Topology*. Springer, New York, corr. 5th print. edition.

Ivanov, A., Nosovskiy, G. V., Chekunov, A. Y., Fedoseev, D. A., Kibkalo, V. A., Nikulin, M., Popelenskiy, F., Komkov, S. A., Mazurenko, I. L., and Petiushko, A. (2021). Manifold hypothesis in data analysis: Double geometrically-probabilistic approach to manifold dimension estimation. *ArXiv*, abs/2107.03903.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*.

Izenman, A. J. (2012). Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):439–446.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Kadanoff, L. P. (1966). Scaling laws for ising models near T_c. *Physics Physique Fizika*, 2:263–272.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

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

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. `http://yann.lecun.com/exdb/mnist/`.

Machta, B. B., Chachra, R., Transtrum, M. K., and Sethna, J. P. (2013). Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607.

McInnes, L. et al. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Meilă, M. and Zhang, H. (2023). Manifold learning: what, how, and why.

Miller, M. (2017). The structure and interpretation of quantum field theory.

Monsalve-Bravo, G. M., Lawson, B. A. J., Drovandi, C., Burrage, K., Brown, K. S., Baker, C. M., Vollert, S. A., Mengersen, K., McDonald-Madden, E., and Adams, M. P. (2022). Analysis of sloppiness in model simulations: Unveiling parameter uncertainty when mathematical models are fitted to data. *Science Advances*, 8(38):eabm5952.

Morrison, M. (2012). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nagel, E. (1961). *The structure of science*. London: Routledge and Keagan Paul.

Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Olah, C. (2014). Neural networks, manifolds, and topology. `https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`. Accessed: 2023-31-12.

Pain, H. J. (2005). *The Physics of Vibrations and Waves*. Wiley, 6 edition.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Polchinski, J. (1984). Renormalization and Effective Lagrangians. *Nucl. Phys. B*, 231:269–295.

Quinn, K. N., Abbott, M. C., Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2022). Information geometry for multiparameter models: new perspectives on the origin of simplicity. *Reports on Progress in Physics*, 86(3):035901.

Raju, A., Machta, B. B., and Sethna, J. P. (2018). Information loss under coarse graining: A geometric approach. *Physical Review E*, 98(5).

Rivat, S. (2021). Effective theories and infinite idealizations: A challenge for scientific realism. *Synthese*.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Ruetsche, L. (2018). Renormalization group realism: The ascent of pessimism. *Philosophy of Science*, 85(5):1176–1189.

Schaffner, K. F. (1967). Approaches to reduction. *Philosophy of Science*, 34:137–147.

Sozou, P., Lane, P., Addis, M., and Gobet, F. (2017). Computational scientific discovery. In Magnani, L. and Bertolotti, T., editors, *Springer Handbook of Model-Based Science*. Springer, Cham.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Stanford, P. K. (2010). *Exceeding Our Grasp*. Oxford University Press USA.

Strandkvist, C., Chvykov, P., and Tikhonov, M. (2020). Beyond rg: from parameter flow to metric flow. *arXiv: Statistical Mechanics*.

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 73(1):109–133.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*.

Teoh, H. K., Quinn, K. N., Kent-Dobias, J., Clement, C. B., Xu, Q., and Sethna, J. P. (2020). Visualizing probabilistic models in minkowski space with intensive symmetrized kullback-leibler embedding. *Phys. Rev. Res.*, 2:033221.

Transtrum, M., Machta, B., Brown, K., Daniels, B., Myers, R., and Sehtna, J. (2015). Sloppiness and emergent theories in physics, biology and beyond. *J. Chem. Phys.*

Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2010). Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.*, 104:060201.

Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2011). Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys. Rev. E*, 83:036701.

Transtrum, M. K. and Qiu, P. (2014). Model reduction by manifold boundaries. *Phys. Rev. Lett.*, 113:098701.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Wallace, D. (2006). In defence of naiveté: The conceptual status of lagrangian quantum field theory. *Synthese*, 151(1):33–80.

Weinberg, S. (1996). What is quantum field theory, and what did we think it is? In *Conference on Historical Examination and Philosophical Reflections on the Foundations of Quantum Field Theory*, pages 241–251.

Williams, P. (2019). Scientific realism made effective. *British Journal for the Philosophy of Science*, 70(1):209–237.

Williamson, J. (2009). The philosophy of science and its relation to machine learning. In Gaber, M., editor, *Scientific Data Mining and Knowledge Discovery*. Springer, Berlin, Heidelberg.

Wilson, K. G. (1983). The renormalization group and critical phenomena. *Rev. Mod. Phys.*, 55:583–600.

Wilson, K. G. and Kogut, J. (1974). The renormalization group and the e expansion. *Physics Reports*, 12(2):75 – 199.

Yao, C., Liu, Y.-F., Jiang, B., Han, J., and Han, J. (2017). Lle score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition. *IEEE Transactions on Image Processing*, 26(11):5257–5269. Epub 2017 Jul 28.

You, S. and Ma, H. (2011). Manifold topological multi-resolution analysis method. *Pattern Recognition*, 44(8):1629–1648.

Zomolodchikov, A. B. (1986). Irreversibility of the Flux of the Renormalization Group in a 2D Field Theory. *JETP Lett.*, 43:730–732.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.