

Article

Solvable Model for the Linear Separability of Structured Data

Marco Gherardi ^{1,2} 

¹ Department of Physics, Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy; marco.gherardi@mi.infn.it

² Istituto Nazionale di Fisica Nucleare Sezione di Milano, via Celoria 16, 20133 Milano, Italy

Abstract: Linear separability, a core concept in supervised machine learning, refers to whether the labels of a data set can be captured by the simplest possible machine: a linear classifier. In order to quantify linear separability beyond this single bit of information, one needs models of data structure parameterized by interpretable quantities, and tractable analytically. Here, I address one class of models with these properties, and show how a combinatorial method allows for the computation, in a mean field approximation, of two useful descriptors of linear separability, one of which is closely related to the popular concept of storage capacity. I motivate the need for multiple metrics by quantifying linear separability in a simple synthetic data set with controlled correlations between the points and their labels, as well as in the benchmark data set MNIST, where the capacity alone paints an incomplete picture. The analytical results indicate a high degree of “universality”, or robustness with respect to the microscopic parameters controlling data structure.

Keywords: linear separability; storage capacity; data structure



Citation: Gherardi, M. Solvable Model for the Linear Separability of Structured Data. *Entropy* **2021**, *23*, 305. <https://doi.org/10.3390/e23030305>

Academic Editor: David Cuesta-Frau

Received: 2 February 2021

Accepted: 25 February 2021

Published: 4 March 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Linear classifiers are quintessential models of supervised machine learning. Despite their simplicity, or possibly because of it, they are ubiquitous: they are building blocks of more complex architectures, for instance, in deep learning and support vector machines, and they provide testing grounds of new tools and ideas in learning theory and statistical mechanics, in both the study of artificial neural networks and in neuroscience [1–9]. Recently, interest in linear classifiers was rekindled by two outstanding results. First, deep neural networks with wide layers can be well approximated by linear models acting on a well defined feature space, given by what is called “neural tangent kernel” [10,11]. Second, it was discovered that deep linear networks, albeit identical to linear classifiers for what concerns the class of realizable functions, allow it to reproduce and explain complex features of nonlinear learning and gradient flow [12].

In spite of the central role that linear separability plays in our understanding of machine learning, fundamental questions still remain open, notably regarding the predictors of separability in real data sets [13]. How does data complexity affect the performance of linear classifiers? Data sets in supervised machine learning are usually not linearly separable: the relations between the data points and their labels cannot be expressed as linear constraints. The first layers in deep learning architectures learn to perform transformations that enhance the linear separability of the data, thus providing downstream fully-connected layers with data points that are more adapted for linear readout [14,15]. The role of “data structure” in machine learning is a hot topic, involving computer scientists and statistical physicists, and impacting both applications and fundamental research in the field [16–22].

Before attempting to assess the effects of data specificities on models and algorithms of machine learning, and, in particular, on the simple case of linear classification, one should have available (i) a quantitative notion of linear separability and (ii) interpretable parameterized models of data structure. Recent advances, especially within statistical mechanics, mainly focused on point (ii). Different models of structured data have been

introduced to express different properties that are deemed to be relevant. For example, the organization of data as the superposition of elementary features (a well-studied trait of empirical data across different disciplines [23–25]) leads to the emergence of a hierarchy in the architecture of Hopfield models [26]. Another example is the “hidden manifold model”, whereby a latent low-dimensional representation of the data is used to generate both the data points and their labels in a way that introduces nontrivial dependence between them [19]. An important class of models assumes that data points are samples of probability distributions that are supported on extended object manifold, which represent all possible variations of an input that should have no effect on its classification (e.g., differences in brightness of a photo, differences in aspect ratio of a handwritten digit) [27]. Recently, a useful parameterization of object manifolds was introduced that is amenable to analytical computations [28]; it will be described in detail below. In a data science perspective, these approaches are motivated by the empirical observation that data sets usually lie on low-dimensional manifolds, whose “intrinsic dimension” is a measure of the number of latent degrees of freedom [29–31].

The main aims of this article are two: (i) the discussion of a quantitative measure of linear separability that could be applied to empirical data and generative models alike; and, (ii) the definition of useful models expressing nontrivial data structure, and the analytical computation, within these models, of compact metrics of linear separability. Most works concerned with data structure and object manifolds (in particular, Refs. [8,27,28]) focus on a single descriptor of linear separability, namely the storage capacity α_c . Informally, the storage capacity measures the maximum number of points that a classifier can reliably classify; in statistical mechanics, it signals the transition, in the thermodynamic limit, between the SAT and UNSAT phases of the random satisfiability problem related to the linear separability of random data [32]. Here, I will present a more complete description of separability than the sole storage capacity (a further motivation is the discovery, within the same model of data structure, of other phenomena lying “beyond the storage capacity” [33]).

2. Linear Classification of Data

Let us first review the standard definition of linear separability for a given data set. In supervised learning, data are given in the form of pairs (ξ_μ, σ_μ) , where $\xi_\mu \in \mathbb{R}^n$ is a data point and $\sigma_\mu = \pm 1$ is a binary label. We focus on dichotomies, i.e., classifications of the data into two subsets (hence, the binary labels); of course, this choice does not exclude datasets with multiple classes of objects, as one can always consider the classification of one particular class versus all the other classes. Given a set of points $X = \{\xi_\mu\}_{\mu=1,\dots,m}$, a dichotomy is a function $\phi : X \rightarrow \{-1, +1\}^m$. A data set $\{(\xi_\mu, \sigma_\mu)\}_{\mu=1,\dots,m}$ is linearly separable (or equivalently the dichotomy $\phi(\xi_\mu) = \sigma_\mu, \mu = 1, \dots, m$, is linearly realizable) if there exists a vector $w \in \mathbb{R}^n$, such that

$$\text{sgn}\left(\sum_{i=1}^n w_i \cdot (\xi_\mu)_i\right) = \sigma_\mu, \quad \mu = 1, \dots, m, \quad (1)$$

where $(\xi_\mu)_i$ is the i th component of the μ th element of the set. In the following, I will simply write $w \cdot \xi_\mu$ for the scalar product appearing in the sgn function when it is obvious that w and ξ_μ are vectors.

In machine learning, the left hand side of Equation (1) is the definition of a linear classifier, or perceptron. The points x , such that $w \cdot x = 0$ define a hyperplane, which is the separating surface, i.e., the boundary between points that are assigned different labels by the perceptron. By viewing the perceptron as a neural network, the vector w is the collection of the synaptic weights. “Learning” in this context refers to the process of adjusting the weight vector w so as to satisfy the m constraints in Equation (1). Because of the fact that the sgn function is invariant under multiplication of its argument by a positive

constant, I will always consider normalized vectors, i.e., both the weight vector w and data points ξ will lie on the unit sphere.

A major motivation behind the introduction of the concept of data structure and the combinatorial theory that is related to it (reviewed in Sections 5 and 6 below) is the fact that the definition of linear separability above is not very powerful per se. Empirically relevant data sets are usually not linearly separable. Knowing whether a data set is linearly separable does not convey much information on its structure: crucially, it does not allow quantifying “how close” to being separable or nonseparable the data set really is. To fix the ideas, let us consider a concrete case: the data set MNIST [34]. MNIST is a collection of handwritten digits, digitized as 28×28 greyscale images, each labelled by the corresponding digit (“0” to “9”). I will use the “training” subset of MNIST, containing 6000 images per digit. To simplify the discussion, I will mainly focus on a single dichotomy within MNIST: that expressed by the labels “3” and “7”. The particular choice of digits is unimportant for this discussion; I will give an example of another dichotomy below, when subtle differences between the digits can be observed.

One may ask the question as to whether the MNIST training set, as a whole, is linearly separable. However, the answer is not particularly informative: the MNIST training set is not linearly separable [34]. But how unexpected is this answer? Can we measure the surprise of finding out a given training set is or is not linearly separable? Intuitively, there are three different properties of a data set that facilitate or hinder its linear separability: size, dimensionality, and structure.

- **Size.** The number of elements m of a data set is a simple indication of its complexity. While a few data points are likely linearly separable, they convey little information on the “ground truth”, the underlying process that generated the data set. On the contrary, larger data sets are more difficult to classify, but the information that is stored in the weights after learning is expected to be more faithful to the ground truth (this is related to the concept of “sample complexity” in machine learning [35]).
- **Dimensionality.** There are two complementary aspects when considering dimensionality in a data oriented framework. First, the embedding dimension is the number of variables that a single data point comprises. For instance, MNIST points are embedded in $\mathbb{R}^{784}$, i.e., each of them is represented by 784 real numbers. The embedding dimension is n in Equation (1); therefore, n is also the number of degrees of freedom that a linear classifier can adjust to find a separating hyperplane. Hence, one expects that a large embedding dimension promotes linear separability. Second, the data set itself does not usually uniformly occupy the embedding space. Rather, points lie on a lower-dimensional manifold, whose dimension d is called the intrinsic dimension of the data set. The concept of general position discussed below is related to the intrinsic dimension; however, beyond that, I will not explicitly consider this type of data complexity in this article (for analytical results on the linear separability of manifolds of varying intrinsic dimension, see [27]).
- **Structure.** As I will show in a moment, the effects of size and dimensionality on linear separability are easily quantified in a simple null model. Data structure, on the other hand, has proved more challenging, and it is the main focus of the theory described here. There is no single definition of data structure; different definitions are useful in different contexts. A common characterization can be given like this: data have structure whenever the data points ξ_μ and their labels σ_μ are not independent variables. I will specify a more precise definition in Section 5. Intuitively, the data structure can both promote or preclude linear separability. If points that are close to one another tend to have the same label then linear separability is improved; if, instead, there are many differently labeled points in a small region of space, then linear separability is obstructed.

Let us get back to the question “how surprising is it that MNIST is not linearly separable?”. This question should be answered by at least taking into account the first two properties described above, the size of the data set and its dimensionality, which are readily computed from the raw data. In fact, the surprise, i.e., the divergence from what is expected based on size and dimensionality, may be interpreted as a beacon of the third property: data structure. I will show in the next section that the answer to our question is “exceedingly unsurprising”. Yet, a slightly modified question will reveal that MNIST, albeit unremarkable in it not being linearly separable, is exceptionally structured.

3. Null Model of Linear Separability

Let us consider a null model of data that fixes the dimension n and the size p . I use a different letter (p instead of m), because it will be useful below to have two different symbols for the size of the whole data set (m) and for the size of its subsets. Consider a data set $Z_p = \{(\xi_\mu, \sigma_\mu)\}_{\mu=1,\dots,p}$, where the vectors ξ_μ are random independent variables that are uniformly distributed on the unit sphere, and the labels σ_μ are independent Bernoulli random variables (also independent from every ξ_μ). These choices are suggested by a maximum entropy principle, when only the parameters m and n are fixed. What is the probability that a data set generated by this model is linearly separable? This problem was addressed and solved more than half a century ago [36–38]; In Section 6 I will describe an analytical technique that allows this computation. The fraction of dichotomies of a random data set that are linearly realizable is

$$c_{n,p} = 2^{1-p} \sum_{i=0}^{n-1} \binom{p-1}{i}, \quad (2)$$

where $\binom{\cdot}{\cdot}$ is the binomial coefficient. Thus, a random (uniform) dichotomy has probability $c_{n,p}$ of being linearly realizable. In this article, I will refer to the probability $c_{n,p}$ as the separability, or probability of separation. A related quantity is the number of dichotomies $C_{n,p} = 2^p c_{n,p}$ (here, 2^p is the total number of dichotomies of p points).

Figure 1 shows the sigmoidal shape of $c_{n,p}$ as a function of p at fixed n . The separability is exactly equal to 1 up to $p = n$ (which pinpoints what is known as the Vapnik–Chervonenkis dimension in statistical learning theory [35]), and it stays close to 1 up to a critical value p_c , which increases with n . At p_c , the curve steeply drops to asymptotically vanishing values, the more abruptly the larger is n . Rescaling the number of points p with the dimension n yields the load $\alpha = p/n$. As a function of α , the probability of separation has the remarkable property of being equal to 1/2 at the critical value (that is known as the storage capacity) $\alpha_c = p_c/n = 2$, independently of n . Such an absence of finite size corrections to the location of the critical point is an unusual feature, which will be lost when we consider structured data below. In the large- n limit, $c_{n,\alpha n}$ converges to a step function that transitions from 1 to 0 at α_c .

How large is the probability of separation $c_{n,m}$ that is given by Equation (2) when one substitutes the sample size $m = 12,000$ and the dimensionality $n = 784$, i.e., those of the dichotomy “3”/“7” in the data set MNIST? The probability, as anticipated, is utterly small, less than 10^{-2000} : it should be no surprise that MNIST is not linearly separable. This comparison is not completely fair, because of the assumption, underlying Equation (2), of general position. The concept of general position is an extension of that of linear independence, which is useful for sets larger than the dimension of the vector space. A set X of vectors in $\mathbb{R}^n$ is in a general position if there is no linearly dependent subset $X' \subseteq X$ of cardinality less than or equal to n . MNIST is quite possibly not in general position. To make sure that it is, I downsampled each image to 10×10 pixels and only considered 1000 images per class (to allow for faster numerical computations), and applied mild multiplicative random noise, by flipping 5% of the pixels around the middle grey value (see Figure 2); I will refer to this modified dataset as “rescaled MNIST”. Running the standard perceptron algorithm on rescaled MNIST did not show signs of convergence

after 10^5 iterations, which indicated that the data set is likely not linearly separable. For $m = 2000$ and $n = 100$, the separability $c_{n,m}$ is less than 10^{-400} .

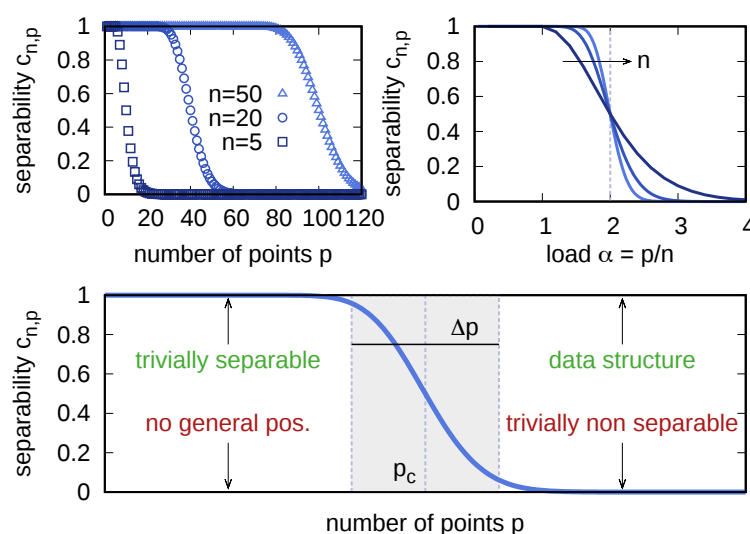


Figure 1. (Top left) The probability of separation, Equation (2), as a function of the number of points p for three values of the embedding dimension n . (Top right) As a function of the load $\alpha = p/n$, the probability of separation converges, for large n , to a step function. (Bottom) Depending on the values of n and p , a data set being separable or nonseparable can convey information about its structure. The location p_c and the width Δp of the transition region are the two main descriptors of the shape of a separability curve.

The null model provides a simple concise interpretation of the linear separability of a given data set, given its size m and dimensionality n , in terms of 5 possible outcomes (see Figure 1, bottom panel):

1. The set is linearly separable and it lies in the region where $c_{n,m} \approx 1$. Separability here is trivial: almost all data sets are separable in this region, provided that the points are in general position.
2. The set is not linearly separable and it lies in the region where $c_{n,m} \approx 1$. The only way this can happen for $m \leq n$ is if the points are not in a general position. For $m > n$, but still in this region, the lack of separability could also be attributed to a non-trivial data structure.
3. The set is not linearly separable and it lies in the region where $c_{n,m} \approx 0$. Almost no dichotomy is linearly realizable in this region; therefore, the lack of separability is trivial here.
4. The set is linearly separable and it lies in the region where $c_{n,m} \approx 0$. This situation is the hallmark of data structure. The fact that the data set happens to represent one of the few dichotomies that are linearly realizable in this region indicates a non-null dependence between the labels and the points in the data set.
5. The set lies in the region where $c_{n,m}$ is significantly different from 0 and 1. Here, knowing that a data set is linearly separable or not is unsurprising either way. The location and the width of this “transition region” are the two main parameters that summarize the shape of the separability curve. In Section 6 I will show how to compute these quantities within a more general model that includes data structure.

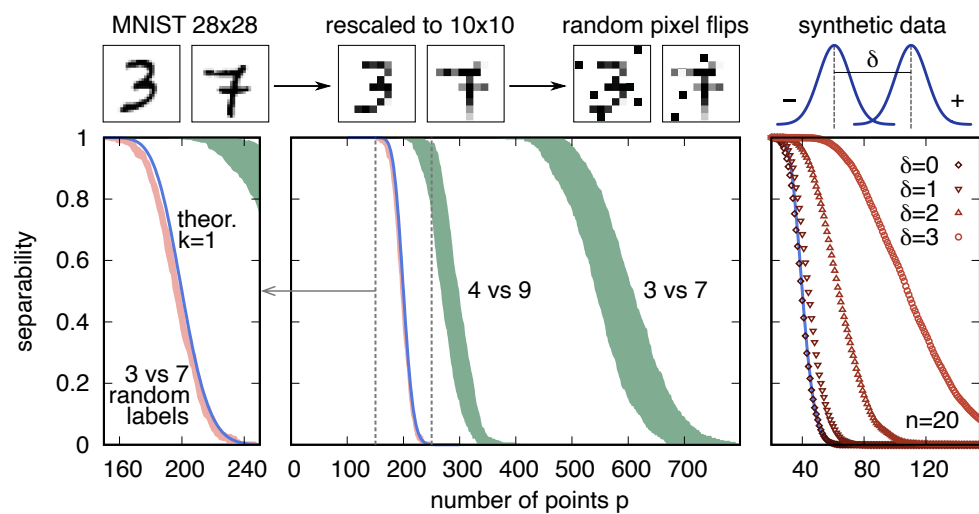


Figure 2. Linear separability (y axis) for subsets of varying size p (x axis), computed in a modified MNIST data set, generated by downscaling and applying multiplicative noise (**left** and **center** panels), and in synthetic data sets generated from a mixture of two normal distributions (**right** panel). (**Left** panel) If the labels are reshuffled, MNIST data (pink area) almost perfectly follow the prediction of the null model (blue line). (**Center** panel) The separabilities of two representative dichotomies in the data set (digits “4” versus “9”, and digits “3” versus “7”) are far removed from the null model, as is apparent from the location (and the width) of their transition regions (green areas). The shaded areas denote the 95% variability intervals. (**Right** panel) By increasing the distance δ between the means of the two Gaussian distributions that define the synthetic data set (here in $n = 20$ dimensions), the separability increases. For $\delta = 0$ (squares), one recovers the prediction of the null model (blue line). Error bars (not shown) are approximately the same size as the symbols.

4. Quantifying Linear Separability via Relative Entropy

In order to make a step further in the characterization of the linear separability of (rescaled) MNIST, we can consider its subsets. While there is only one subset with $m = 2000$ points (focusing on the dichotomy “3”/“7”), and only one yes/no answer to the question of its linear separability, there are many subsets of size $p < m$, which can provide more detailed information. To quantify such information, let us formulate a more precise notion of surprise with respect to a model expressing prior expectation [39]. Let us again fix an empirical data set $Z_m = \{(\xi_\mu, \sigma_\mu)\}_{\mu=1, \dots, m}$ and fix $p \leq m$. Now, consider the set N_p of all subsets $\nu = \{v_1, \dots, v_p\}$ of p indices $v_i \in \{1, \dots, m\}$, with $v_i \neq v_j$ for $i \neq j$. Additionally, consider the set $\Sigma_p = \{-1, +1\}^p$ of all dichotomies $\hat{\sigma} = \{\hat{\sigma}_1, \dots, \hat{\sigma}_p\}$ of p elements. (I use curly braces for both sets and indexed families.) For each pair $(\nu \in N_p, \hat{\sigma} \in \Sigma_p)$, we can construct the corresponding synthetic dataset

$$Z(\nu, \hat{\sigma}) = \{(\xi_{v_i}, \hat{\sigma}_i)\}_{i=1, \dots, p} \quad (3)$$

similarly, for each $\nu \in N_p$, we can construct the corresponding subset $Z_{\text{emp}}(\nu)$ of the empirical data set Z_m :

$$Z_{\text{emp}}(\nu) = \{(\xi_{v_i}, \sigma_{v_i})\}_{i=1, \dots, p}. \quad (4)$$

The main tool for defining the surprise will be probability distributions on a space Ω_p , which is defined as the union of all synthetic data sets:

$$\Omega_p = \bigcup_{(\nu, \hat{\sigma}) \in N_p \times \Sigma_p} Z(\nu, \hat{\sigma}). \quad (5)$$

The empirical space $\Omega_p^{\text{emp}} \subseteq \Omega_p$ can be defined similarly:

$$\Omega_p^{\text{emp}} = \bigcup_{\nu \in N_p} Z_{\text{emp}}(\nu). \quad (6)$$

Essentially, Ω_p^{emp} contains all collections of p point/label pairs in the data set Z_m , while Ω_p contains all the collections of p point/label pairs where the p points are chosen among the ones in the data set and the labels are all possible 2^p combinations on those p points. Notice that Ω_p and Ω_p^{emp} have different cardinalities: $|\Omega_p^{\text{emp}}| = M_p$ and $|\Omega_p| = 2^p M_p$, where $M_p = \binom{m}{p}$ is the number of subsets of size p in the data set.

Interpreted as a probability distribution on Ω_p , the empirical data are uniform distributed on Ω_p^{emp} ; likewise, the null model defined above induces, by conditioning on the points $\{\xi_\mu\}$, the uniform distribution on the whole Ω_p . In general, not every data set in Ω_p (nor in Ω_p^{emp}) is linearly separable. Let us define the subsets for which this property holds:

$$\begin{aligned} \hat{\Omega}_p &= \{Z \in \Omega_p : Z \text{ is linearly separable}\} \\ \hat{\Omega}_p^{\text{emp}} &= \{Z \in \Omega_p^{\text{emp}} : Z \text{ is linearly separable}\}. \end{aligned} \quad (7)$$

Let us call Q_p and Q_p^{emp} the uniform probability distributions on $\hat{\Omega}_p$ and $\hat{\Omega}_p^{\text{emp}}$, respectively. The Kullback–Leibler (KL) divergence $D_{\text{KL}}(Q_p^{\text{emp}} \| Q_p)$ from Q_p to Q_p^{emp} (or relative entropy)

$$D_{\text{KL}}(Q_p^{\text{emp}} \| Q_p) = \sum_{z \in \Omega_p} Q_p^{\text{emp}} \log \frac{Q_p^{\text{emp}}}{Q_p} \quad (8)$$

then measures the surprise carried by the data with respect to the prior belief regarding its linear separability expressed by Q_p . Because Q_p and Q_p^{emp} are defined on sets $(\Omega_p$ and $\Omega_p^{\text{emp}})$ of different cardinality, I define the (signed) surprise $\mathcal{S}_p$ by subtracting the reference KL divergence between the uniform distributions on these spaces:

$$\begin{aligned} \mathcal{S}_p &= D_{\text{KL}}(Q_p^{\text{emp}} \| Q_p) - \log(|\Omega_p| / |\Omega_p^{\text{emp}}|) \\ &= D_{\text{KL}}(Q_p^{\text{emp}} \| Q_p) - p \log 2. \end{aligned} \quad (9)$$

Notice that the summand in the definition of KL divergence, Equation (8), is only nonzero for $z \in \hat{\Omega}_p^{\text{emp}}$; one then obtains

$$\begin{aligned} \mathcal{S}_p &= \sum_{z \in \hat{\Omega}_p^{\text{emp}}} \frac{1}{|\hat{\Omega}_p^{\text{emp}}|} \log \frac{|\hat{\Omega}_p|}{|\hat{\Omega}_p^{\text{emp}}|} - p \log 2 \\ &= \log \frac{|\hat{\Omega}_p|}{|\hat{\Omega}_p^{\text{emp}}|} - p \log 2 \\ &= \log \frac{c_{n,p} M_p 2^p}{|\hat{\Omega}_p^{\text{emp}}|} - p \log 2 \\ &= \log c_{n,p} - \log c_{n,p}^{\text{emp}}, \end{aligned} \quad (10)$$

where I have defined the empirical separability $c_{n,p}^{\text{emp}}$ as the fraction of linearly separable subsets of size p in Z_m :

$$c_{n,p}^{\text{emp}} = \frac{|\hat{\Omega}_p^{\text{emp}}|}{M_p}. \quad (11)$$

The signed surprise $\mathcal{S}_p$ is positive (respectively negative) when the fraction of linearly separable subsets of size p is smaller (respectively larger) than expected in the null model.

Separability in a Synthetic Data Set and in MNIST

The discussion above encourages the use of the empirical separability $c_{n,p}^{\text{emp}}$ as a detailed description of the linear separability of a data set in an information theoretic framework. Despite being one of the simplest benchmark data sets used in machine learning, MNIST is already rather complex; its classes are known to have small intrinsic dimensions and varied geometries [15]. Therefore, before turning to MNIST, let us consider a simple controlled experiment, where the data are extracted from a simple one-parameter mixture distribution, defined, as follows. Let $\sigma \in \{-1, +1\}$ be a Bernoulli random variable with parameter $1/2$, which generates the labels. The data points $\xi \in \mathbb{R}^n$ are extracted from a multivariate normal distribution with σ -dependent mean. The joint probability distribution of each point-label pair is

$$P(\{\xi, \sigma\}) = \frac{1}{2} f_{\mathcal{N}(\mu(\sigma), \mathbb{I})}(\xi), \quad \mu(\sigma) = \left(\sigma \frac{\delta}{2}, 0, \dots, 0 \right), \quad (12)$$

where $f_{\mathcal{N}(\mu, \mathbb{I})}$ is the probability density function of the multivariate normal distribution with mean μ and identity covariance matrix. The parameter δ measures the distance between the two means: $\delta = \|\mu(\sigma = +1) - \mu(\sigma = -1)\|$. Figure 2 shows the empirical separability $c_{n,p}^{\text{emp}}$, as a function of the size p of the subsets, for such a data set containing $m = 200$ data points in $n = 20$ dimensions. When $\delta = 0$, all of the data points are extracted from the same distribution, regardless of their labels: the data have no structure and the separability follows the null model, as in Equation (2). While δ increases, equally labelled points start to cluster, and the separability at any given $p > n$ increases, as expected from the qualitative discussion in Section 2. It is interesting to note that the width of the transition region (Δp in Figure 1) is also an increasing function of δ . This dependence was not expected *a priori*; In Section 7, I will show that the theory of structured data presented below allows for explaining this behavior.

Let us now compute $c_{n,p}^{\text{emp}}$ for the rescaled MNIST data set. Figure 2 shows the results of three numerical experiments, as compared with the null model prediction (2), and elicits four observations. (i) MNIST data are significantly more separable than the null model. For instance, the signed surprise, with respect to the null model, of the empirical dichotomies separating the digits “3” and “7” takes the values $S_{400} \approx -55$, $S_{500} \approx -100$, $S_{600} \approx -150$. (ii) Even within the same data set, different classifications can have different probabilities of separation; the dichotomy separating the digits “4” and “9” in rescaled MNIST is closer to the null model than the dichotomy of “3” and “7” (e.g., $S_{400} \approx -48$). (iii) Destroying the structure by random reshuffling of the labels makes the separability collapse onto that of the null model; the surprise S_p in this case is, at most, of order 10^{-1} for all p . (iv) Similarly to what happens in the more controlled experiment with the synthetic data above, the separability curve of the “3”/“7” dichotomy, which has its transition point at a larger value of p than the “3”/“9” dichotomy, also has a wider transition region.

This analysis shows that, contrary to what appeared by looking solely at the whole data set, the dichotomies of rescaled MNIST are much more likely to be realized by a linear separator than random ones. In relation to the separability as a function of p , the null model has a single parameter, the dimension n . Is it possible to interpret the empirical curves as those of the null model with an effective dimension n_{eff} ? Increasing n has the effect of increasing proportionally the value p_c because the storage capacity is fixed to $\alpha_c = 2$. However, while fixing $n_{\text{eff}} \approx 280$ indeed aligns the critical number of points p_c with the empirical one, it yields a much smaller width of the transition region ($\Delta p \approx 80$ for the null model and $\Delta p \approx 300$ in the data). Furthermore, notice that the values of the surprise for the “3”-vs.-“7” and “4”-vs.-“9” experiments are not very different. The reason is the ingenuousness of the null model, which hardly captures the properties of the empirical sets, and whose term $c_{n,p}$ therefore dominates in S_p . These observations, together with the motivations that are discussed above, are a spur for the definition of a more nuanced and versatile model of the separability of structured data.

5. Parameterized Model of Structured Data

Fixing a model of data structure in this context means fixing a generative model of data. Here, I use the model first introduced in [28]. This should not be considered to be a realistic model of real data sets. It is useful as an effective or phenomenological parameterization of data structure. It has two main advantages: (i) it allows the analytical computation, within a mean field approximation, of the probability of separation $c_{n,p}$; and, (ii) it naturally points out the relevant geometric-probabilistic parameters that control the linear separability.

The model is expressed in the form of constraints between the points and the labels. The synthetic data set is constructed as a collection of q “multiplets”, i.e., subsets of k points $\{\xi_\mu^1, \dots, \xi_\mu^k\}$ with prescribed geometric relations between them, and such that the labels are constant within each multiplet:

$$Z_q = \left\{ \begin{aligned} &(\xi_1^1, \sigma_1), \dots, (\xi_1^k, \sigma_1), \\ &(\xi_2^1, \sigma_2), \dots, (\xi_2^k, \sigma_2), \\ &\vdots \\ &(\xi_q^1, \sigma_q), \dots, (\xi_q^k, \sigma_q) \end{aligned} \right\}. \quad (13)$$

The total number of point/label pairs is $p = qk$. Observe that, if one considers the set of all points $X = \{\xi_\mu^i\}$, not every dichotomy of X is admitted by the parameterization of Z_q in Equation (13). If a dichotomy assigns different labels to two elements of the same multiplet, it cannot be written in this form. The dichotomies that agree with the parameterization of Equation (13) are termed as admissible.

The relations between the points ξ_μ^i within each multiplet can be fixed, for instance, by prescribing that the $k(k-1)/2$ overlaps $\rho_{i,j} = \xi_\mu^i \cdot \xi_\mu^j$ be fixed and independent of μ (remember that $|\xi_\mu^i| = 1$). The statistical ensemble for Z_q , as specified by the probability density $dp(Z_q)$, is chosen in accordance with the maximum entropy principle: it is the uniform probability distribution on the points and the labels independently, given the constraints:

$$dp(Z_q) = \frac{1}{\mathcal{Z}(n, q, \{\rho_{i,j}\})} \prod_{\substack{\mu=1, \dots, q \\ i=1, \dots, k}} d\xi_\mu^i \delta(|\xi_\mu^i| - 1) \prod_{a>b=1, \dots, k} \delta(\rho_{a,b} - \xi_\mu^a \cdot \xi_\mu^b), \quad (14)$$

where $\mathcal{Z}(n, q, \{\rho_{i,j}\})$ is the partition function, fixed by the normalization condition

$$\sum_{\{\sigma_\mu\}} \int_{\mathbb{R}^{nqk}} dp(Z_q) = 1. \quad (15)$$

The null (unstructured) model of Section 3 is recovered in this parameterization in two different limits. First, if $k = 1$ each multiplet is composed of a single point, and no constraints are imposed other than the normalization. Second, for any k , if all overlaps are fixed to 1, then all points in each overlap coincide, $\xi_\mu^1 = \xi_\mu^2 = \dots = \xi_\mu^k$, and the model is equivalent to the null model with $p = q$.

The theory that will be described below depends on a natural set of parameters ψ_m , with $m = 2, \dots, k$. These quantities are conditional probabilities of geometric events that are related to single multiplets. They characterize the properties of the multiplets that are relevant for the linear separability of the whole set. Consider a multiplet $X = \{\xi^1, \dots, \xi^k\}$. ψ_m is a measure of the likelihood that a subset $X' \subseteq X$ of $m \leq k$ points is classified coherently by a random weight vector. More precisely, ψ_m is the probability that the scalar product $w \cdot \xi$ has the same sign for all $\xi \in X'$, being conditioned on the event that $w \cdot \xi$ has the same sign for all $\xi \in X \setminus \{\xi_\star\}$. This probability is computed in the ensemble where the

vector w is uniformly distributed on the unit sphere S^{n-1} , X' is uniformly distributed on the subsets of X of m points, and ξ_* is uniformly distributed on the elements of X' . This is coherent with the mean field nature of the combinatorial theory, which assumes uniformly distributed and uncorrelated quantities (see below).

In a few cases, ψ_m can be computed explicitly. For instance, for a doublet $\{\xi, \bar{\xi}\}$ at fixed overlap $\rho = \xi \cdot \bar{\xi}$,

$$\psi_2(\rho) = \frac{2}{\pi} \arctan \sqrt{\frac{1+\rho}{1-\rho}}. \quad (16)$$

This is the probability that a random hyperplane does not intersect the segment that connects two points at overlap ρ . It is an increasing function of ρ , from $\psi_2(-1) = 0$ to $\psi_2(1) = 1$. If $k > 2$, then the quantity that enters the equations will be the mean of $\psi_2(\rho)$ over all the pairs in the multiplet. It can be shown that ψ_m , as a function of the overlaps $\rho_{i,j}$, does not explicitly depend on the dimensionality n [28]; this property greatly simplifies the analytical computations.

In summary, the parameters of the model are the following: the dimensionality n , the multiplicity k , and the $k-2$ probabilities ψ_m . Actually, only two special combinations of the parameters ψ_m emerge as relevant from the theory that is presented in the next sections:

$$\Psi_1 = \sum_{r=2}^k \psi_r, \quad (17)$$

$$\Psi_2 = \sum_{r=2}^k \sum_{l=2}^{r-1} \psi_r \psi_l. \quad (18)$$

I will call them structure parameters. Other functions of the probabilities ψ_m are relevant for other purposes, for instance, when considering the large- p asymptotics of $c_{n,p}$, which relates to the generalization properties of the linear separator [32].

6. Combinatorial Computation of the Separability for Structured Data

Cover popularized a powerful combinatorial technique to compute the number of linearly realizable dichotomies in an old and highly cited paper [38]. Despite its appeal, the combinatorial approach (while certainly not extraneous to contemporary statistical physics, both theoretical and applied [40–43]) remained somewhat confined to very few papers in discrete mathematics, and it was only very recently extended to more modern questions, when it was used to obtain an equation for $C_{n,q}$, the number of admissible dichotomies of q multiplets, for structured data of the type that is defined in the previous section. Ref. [28] first presented the arguments and computations leading to this equation. To make this article as self-contained as possible, I repeat most of the derivation here.

6.1. Exact Approach for Unstructured Data ($k = 1$ Points per Multiplet)

First, I recall the classic computation for unstructured data ($k = 1$ in our notation). The idea is to write a recurrence relation for the number of linearly realizable dichotomies $C_{n,p}$ and, consequently, for the probability $c_{n,p}$, by considering the addition of the $(p+1)$ th element ξ_{p+1} to the set $X_p = \{\xi_1, \dots, \xi_p\}$ that was composed of the first p elements.

Consider one of the dichotomies of X_p , let us call it ϕ_p ; how many linearly realizable dichotomies of $X_{p+1} = \{\xi_1, \dots, \xi_p, \xi_{p+1}\}$ agree with ϕ_p (i.e., take the same values) on the points of X_p ? When the point ξ_{p+1} is added to the set, two different things can happen: (i) $\text{sgn}(w \cdot \xi_{p+1})$ is the same for all possible weight vectors w that realize ϕ_p ; and, (ii) there is at least one weight vector $\hat{w}$ realizing ϕ_p , such that $\hat{w} \cdot \xi_{p+1} = 0$. These two cases lead to different contributions to $C_{n,p+1}$. In the first case, there is only one dichotomy of X_{p+1} agreeing with ϕ_p , as the value that is assigned to ξ_{p+1} is fixed. In the second case, the value that is assigned to ξ_{p+1} can be either $+1$ or -1 ; therefore, the number of dichotomies of X_{p+1} agreeing with ϕ_p is 2.

Let us call $M_{n,p}$ the number of those dichotomies, among the $C_{n,p}$ dichotomies of X_p , such that (ii) holds for the new point; the number of those satisfying (i) will be $C_{n,p} - M_{n,p}$. The reasoning above then leads to $C_{n,p+1} = (C_{n,p} - M_{n,p}) + 2M_{n,p} = C_{n,p} + M_{n,p}$. Here lies the keystone that allows for the closure of the recurrence equation: $M_{n,p}$ is the number of dichotomies conditioned to satisfy a linear constraint; therefore, it is equal to the number of dichotomies, of the same number of points p , in $n - 1$ dimensions: $M_{n,p} = C_{n-1,p}$. Finally, the recurrence relation is $C_{n,p+1} = C_{n,p} + C_{n-1,p}$, which translates into the following equation for the probability $c_{n,p}$:

$$c_{n,p+1} = \frac{1}{2}(c_{n,p} + c_{n-1,p}). \quad (19)$$

The boundary conditions of the recurrence (19) are

$$\begin{aligned} c_{n>0,1} &= 1, \\ c_{n\leq 0,p} &= 0 \quad \left[\Rightarrow c_{1,p} = 2^{1-p} \right], \end{aligned} \quad (20)$$

which come from the conditions $C_{1,p>0} = 2$ (there are only two normalized weight vectors in one dimension) and $C_{n>0,1} = 2$ (there is always a weight vector w , such that $\pm w \cdot \xi = \pm 1$). The solution of Equation (19) is Equation (2), as can be checked directly. However, the more complicated equations that are satisfied by the probabilities for structured data are not as easily solvable. For this reason, in Section 7, below, I will show a method to compute useful quantities that are related to the shape of $c_{n,p}$ directly from the recurrence relations, with no need for a closed solution.

6.2. Mean-Field Approach for Pairs of Points ($k = 2$ Points per Multiplet)

The simplest non-trivial extension of Cover's computation to structured data is $k = 2$. From here on I will use $\hat{c}_{n,q}$ and $\hat{C}_{n,q}$ to denote the fraction and number of linearly realizable admissible dichotomies of q multiplets because the symbols $c_{n,p}$ and $C_{n,p}$ were reserved to denote the fraction and number of linearly realizable dichotomies of p points.

Notice that all the quantities appearing above are notated with no explicit dependence on the points ξ . This is because the unstructured case enjoys a strong universality property (as proved in [38]): $C_{n,p}$ is independent of the points of X_p , as long as they are in a general position. Such generality breaks down for structured data. In this case, the recurrence equations that will be obtained are not valid for all sets X_p ; rather, they are satisfied by the ensemble averages of $\hat{C}_{n,q}$ and $\hat{c}_{n,q}$, in the spirit of the mean-field approximation of statistical physics.

The set of points is now $X_q \cup \bar{X}_q$, where X_q is a set of q points $\{\xi_1, \dots, \xi_q\}$ and $\bar{X}_q$ is a set of partners $\{\bar{\xi}_1, \dots, \bar{\xi}_q\}$, where $\xi_\mu \cdot \bar{\xi}_\mu = \rho$ for all $\mu = 1, \dots, q$ (remember that all of the points are on the unit sphere). Consider the addition of the points ξ_{q+1} and $\bar{\xi}_{q+1}$ to X_q and $\bar{X}_q$, respectively. By repeating the reasoning described above for $k = 1$ with respect to the point $\bar{\xi}_{q+1}$, one finds a formula for the number $Q_{n,q}$ of dichotomies of the set $\{\xi_1, \bar{\xi}_1, \dots, \xi_q, \bar{\xi}_q, \xi_{q+1}, \bar{\xi}_{q+1}\}$ that are admissible on the first q pairs (and are unconstrained on $\bar{\xi}_{q+1}$): $Q_{n,q} = \hat{C}_{n,q} + \hat{C}_{n-1,q}$. These dichotomies can be separated into two classes, similarly to the two cases (i) and (ii) above: those that can be realized by a weight vector orthogonal to ξ_{q+1} (let us denote their number by $R_{n,q}$) and those that cannot (their number is then $Q_{n,q} - R_{n,q}$). For each dichotomy ϕ of the first class, there exists one and only one admissible dichotomy of the full set $X_{q+1} \cup \bar{X}_{q+1}$ that agrees with ϕ and can be realized linearly. In fact, thanks to the orthogonality constraint, there is always, among the weight vectors realizing ϕ , one vector w , such that

$$\text{sgn}(w \cdot \bar{\xi}_{q+1}) = \phi(\bar{\xi}_{q+1}), \quad (21)$$

thus satisfying the admissibility condition on the pair $\{\xi_{q+1}, \bar{\xi}_{q+1}\}$. The remaining $Q_{n,q} - R_{n,q}$ dichotomies do not allow this freedom. How many of them are realized by weight

vectors w , such that the admissibility condition (21) is satisfied can be estimated at the mean field level by the probability that, given a random weight vector w chosen uniformly on the unit sphere, the scalar products $w \cdot \xi_{q+1}$ and $w \cdot \bar{\xi}_{q+1}$ have the same sign. This probability does not depend on the actual points, but only on their overlap ρ , and it is exactly the quantity $\psi_2(\rho)$ that is defined in the previous section, Equation (16). I will denote it by ψ_2 in the following, with the dependence on ρ being understood.

The foregoing argument brings the following equation:

$$\hat{c}_{n,q+1} = R_{n,q} + \psi_2(\hat{c}_{n,q} + \hat{c}_{n-1,q} - R_{n,q}) \quad (22)$$

Similarly to what happens in the unstructured case, the unknown term $R_{n,q}$ can be expressed in terms of variables $\hat{c}_{\bullet,q}$ by considering the same problem in a lower dimension. In fact, remember that $Q_{n,q}$ above was computed by applying Cover's argument for $k = 1$, because it counts how the number of dichotomies is affected when the single point $\bar{\xi}_{q+1}$ is added to the set. $R_{n,q}$ must be computed in the same way, since it, again, counts the number of dichotomies that are admissible on the first q pairs and free on $\bar{\xi}_{q+1}$. However, these dichotomies must satisfy the additional linear constraint $w \cdot \xi_{q+1} = 0$; therefore, the whole argument must be applied in $n - 1$ dimensions. This leads to

$$R_{n,q} = \hat{c}_{n-1,q} + \hat{c}_{n-2,q}. \quad (23)$$

Finally, substituting this expression of $R_{n,q}$ into Equation (22) yields

$$\hat{c}_{n,q+1} = \psi_2 \hat{c}_{n,q} + \hat{c}_{n-1,q} + (1 - \psi_2) \hat{c}_{n-2,q}. \quad (24)$$

As above, this translates to a similar equation for the probability $\hat{c}_{n,q}$:

$$\hat{c}_{n,q+1} = \frac{\psi_2}{2} \hat{c}_{n,q} + \frac{1}{2} \hat{c}_{n-1,q} + \frac{1 - \psi_2}{2} \hat{c}_{n-2,q}. \quad (25)$$

The boundary conditions of this recurrence are slightly different than for $k = 1$. They are discussed in the Appendix A, together with those for the general case.

6.3. General Case Parameterized by k

It is possible to extend the method that is described above to all k . I will only sketch the derivation; the details can be found in [28]. Just as the case $k = 2$ can be treated by making use of the recurrence formula for $k = 1$, the idea here is to construct the case k recursively by using the formula (yet to be found) for $k - 1$, therefore obtaining a recurrence relation in k as well as in n and q . To this aim, the $(q + 1)$ th multiplet $\{\xi_{q+1}^1, \dots, \xi_{q+1}^k\}$ is split into the two subsets $\{\xi_{q+1}^1\}$ and $\bar{\xi}_{q+1} = \{\xi_{q+1}^2, \dots, \xi_{q+1}^k\}$. The formula for $k - 1$ allows for applying the argument to the set $\bar{\xi}_{q+1}$, thus obtaining the number $Q_{n,q}$ of dichotomies of the set $X_q \setminus \{\xi_{q+1}^1\}$ that are admissible on the first q complete multiplets and are admissible on the $(q + 1)$ th incomplete multiplet $\bar{\xi}_{q+1}$. More formally, $Q_{n,q}$ is the number of linearly realizable dichotomies ϕ , such that

$$\begin{aligned} \phi(\xi_\mu^i) &= \phi(\bar{\xi}_\mu^j) \quad i, j = 1, \dots, k; \mu = 1, \dots, q \\ \phi(\xi_{q+1}^i) &= \phi(\bar{\xi}_{q+1}^j) \quad i, j = 2, \dots, k. \end{aligned} \quad (26)$$

Now the argument goes exactly as for the case $k = 2$: some of these $Q_{n,q}$ dichotomies (their number being $R_{n,q}$) can be realized by a weight vector orthogonal to the point ξ_{q+1}^1 ; therefore, each of them contributes a single admissible dichotomy of the whole set X_{q+1} ; the remaining $Q_{n,q} - R_{n,q}$ contribute with probability ψ_k . Again, $R_{n,q}$ can be expressed by applying the same argument in $n - 1$ dimensions.

Finally, one finds that the probability $\hat{c}_{n,q}$ satisfies a recurrence equation in n and q :

$$\hat{c}_{n,q+1} = \sum_{l=0}^k \theta_l^k c_{n-l,q}, \quad (27)$$

where the coefficients θ_l^k are constants (independent of n and q) satisfying a recurrence equation in k and l :

$$\theta_l^k = \psi_k \theta_l^{k-1} + (1 - \psi_k) \theta_{l-1}^{k-1}. \quad (28)$$

The boundary conditions for Equation (28) are

$$\begin{aligned} \theta_0^1 &= \theta_1^1 = \frac{1}{2} \\ \theta_{l < 0}^k &= \theta_{l > k}^k = 0; \end{aligned} \quad (29)$$

the conditions at $k = 1$ are those that reproduce Equation (19).

7. Computation of Compact Metrics of Linear Separability

The model of data structure leading to the foregoing equations is very detailed, in that it allows for the independent specification of a large number of parameters. However, the influence of each parameter on the separability $\hat{c}_{n,q}$ is not equal, with some combinations of parameters being more relevant than others. In this section, I compute two main descriptors of the shape of $\hat{c}_{n,q}$ as a function of q at n fixed: the transition point p_c (equivalently, the capacity α_c) and the width Δp of the transition region; they are defined more precisely below. We will see that only the structure parameters Ψ_1 and Ψ_2 , the special combinations defined in Section 5, are needed to fix p_c and Δp .

7.1. Diagonalization of the Recurrence Relation

Notice that, while the quantity $\hat{c}_{n,q}$ that is given by the theory is expressed as a function of the number of multiplets q , the definition of separability that is discussed in Section 5 is given in terms of the number of points $p = kq$. This is not really a problem in the thermodynamic limit

$$\begin{aligned} n &\rightarrow \infty \\ p &\rightarrow \infty \\ \alpha &= \frac{p}{n} \text{ fixed} \\ k &\text{ fixed,} \end{aligned} \quad (30)$$

whereby the separability is expressed as a function of the load α . In the following, I will define the location q_c and the width Δq of the transition region in the parameterization by the number of multiplets q ; the corresponding quantities that are parameterized by p are obtained by rescaling:

$$p_c = kq_c, \quad \Delta p = k\Delta q. \quad (31)$$

Let us consider the discrete derivative of $\hat{c}_{n,q}$ with respect to n :

$$\gamma_{n,q} = \Delta_n \hat{c}_{n,q} \equiv \hat{c}_{n+1,q} - \hat{c}_{n,q}. \quad (32)$$

As will be clear momentarily, working with $\gamma_{n,q}$ is convenient because it is normalized, as I will prove below. $\gamma_{n,q}$ satisfies the same recurrence relation as $\hat{c}_{n,q}$:

$$\gamma_{n,q+1} = \sum_{l=0}^k \theta_l^k \gamma_{n-l,q}. \quad (33)$$

The boundary conditions, in accordance with (20), are

$$\begin{aligned}\gamma_{n,1} &= \delta_{n,0}, \\ \gamma_{n<0,q} &= 0.\end{aligned}\quad (34)$$

The right hand side of Equation (33) has the form of a discrete convolution between $\theta_{\bullet}^k$ and $\gamma_{\bullet,q}$:

$$c_{\bullet,q+1} = \theta_{\bullet}^k * c_{\bullet,q}. \quad (35)$$

The convolution is diagonalized in Fourier space, by defining the characteristic functions

$$\tilde{\gamma}_q(t) = \sum_{n=0}^{\infty} \gamma_{n,q} e^{int}, \quad (36)$$

$$\tilde{\theta}_k(t) = \sum_{n=0}^{\infty} \theta_n^k e^{int}. \quad (37)$$

Multiplying both sides of Equation (35) by e^{int} and summing over n yields

$$\tilde{\gamma}_{q+1}(t) = \tilde{\theta}_k(t) \tilde{\gamma}_q(t). \quad (38)$$

From the definition (36) and boundary conditions (34), one gets $\tilde{\gamma}_1(t) = 1$; hence, the solution of the recurrence equation is

$$\tilde{\gamma}_q(t) = [\tilde{\theta}_k(t)]^{q-1}. \quad (39)$$

7.2. Defining the Location and Width of the Transition Region

As mentioned above, $\gamma_{n,q}$ is normalized, which means that

$$\sum_{n=0}^{\infty} \gamma_{n,q} = 1, \quad (40)$$

or, equivalently, $\tilde{\gamma}_q(0) = 1$. To prove this, it suffices to show that $\tilde{\theta}_k(0) = 1$, i.e., that θ_n^k is normalized. Summing both sides of Equation (28) in l from 0 to ∞ shows that $\tilde{\theta}_k(0)$ is constant in k , therefore

$$\tilde{\theta}_k(0) = \tilde{\theta}_1(0) = 1, \quad (41)$$

as can be computed from the boundary conditions (29).

Because it is normalized, $\gamma_{\bullet,q}$ can be interpreted as a probability distribution, whose cumulative distribution function is $\hat{c}_{\bullet,q}$. The a th moment of the distribution is

$$\begin{aligned}\langle n^a \rangle_q &= \sum_{n=0}^{\infty} n^a \gamma_{n,q} \\ &= i^{-a} \frac{d^a}{dt^a} \tilde{\gamma}_q(t) \Big|_{t=0}.\end{aligned}\quad (42)$$

The same holds for $\theta_{\bullet}^k$, whose moments $\langle \theta^a \rangle_k$ can be obtained from its characteristic function $\tilde{\theta}_k(t)$. Let us focus on the mean μ_q and the variance σ_q ,

$$\mu_q = \langle n \rangle_q, \quad \sigma_q^2 = \langle n^2 \rangle_q - \langle n \rangle_q^2. \quad (43)$$

Equation (39) allows for expressing these quantities in terms of the mean $\mu_{\theta} = \langle \theta \rangle_k$ and variance $\sigma_{\theta}^2 = \langle \theta^2 \rangle_k - \langle \theta \rangle_k^2$ of $\theta_{\bullet}^k$:

$$\mu_q = (q-1)\mu_{\theta}, \quad \sigma_q^2 = (q-1)\sigma_{\theta}^2, \quad (44)$$

as can be checked by using Equation (42).

We can now define the two main descriptors, q_c and Δq , which summarize the separability as a function of q :

$$q_c : \mu_{q_c} = n, \quad (45)$$

$$q_{\pm} : \mu_{q_{\pm}} \mp \sigma_{q_{\pm}} = n, \quad (46)$$

$$\Delta q : \Delta q = q_+ - q_-. \quad (47)$$

7.3. Expression in Terms of the Structure Parameters

To compute these quantities, all we need is μ_{θ} and σ_{θ} , or $\langle \theta \rangle_k$ and $\langle \theta^2 \rangle_k$. Solving Equation (45) for q_c gives

$$q_c = n\mu_{\theta}^{-1} + 1. \quad (48)$$

Solving Equations (46) and (47) for Δq gives

$$\Delta q = \frac{1}{\mu_{\theta}^2} \sqrt{\sigma_{\theta}^2 (\sigma_{\theta}^2 + 4\mu_{\theta}n)}. \quad (49)$$

The corresponding expressions to leading order in n are the following

$$\begin{aligned} q_c &= n\mu_{\theta}^{-1} + O(1), \\ \Delta q &= 2\sigma_{\theta}\mu_{\theta}^{-3/2}n^{1/2} + O(n^{-1/2}). \end{aligned} \quad (50)$$

The moments of θ_k^* satisfy the following equation, which can be obtained by multiplying both sides of Equation (28) by l^a and summing over l :

$$\begin{aligned} \langle \theta^a \rangle_k &= \psi_k \langle \theta^a \rangle_{k-1} + (1 - \psi_k) \sum_{l=0}^{\infty} (l+1)^a \theta_l^{k-1} \\ &= \langle \theta^a \rangle_{k-1} + (1 - \psi_k) \sum_{s=0}^{a-1} \binom{a}{s} \langle \theta^s \rangle_{k-1}. \end{aligned} \quad (51)$$

The boundary conditions are $\langle \theta^0 \rangle_k = 1$ (computed above) and $\langle \theta^a \rangle_1 = 1/2$, as given by Equation (29). In particular, for $a = 1$, we obtain

$$\langle \theta \rangle_k = \langle \theta \rangle_{k-1} + (1 - \psi_k), \quad (52)$$

whose solution is

$$\langle \theta \rangle_k = k - \frac{1}{2} - \Psi_1, \quad (53)$$

where the structure parameter Ψ_1 , as defined in Equation (17), implicitly depends on k . For $a = 2$, the recurrence Equation (51) becomes

$$\langle \theta^2 \rangle_k = \langle \theta^2 \rangle_{k-1} + (1 - \psi_k) (2\langle \theta \rangle_{k-1} + 1). \quad (54)$$

By substituting $\langle \theta \rangle_{k-1}$ given by Equation (53) and solving the recurrence we obtain, after some algebra,

$$\langle \theta^2 \rangle_k = k^2 - k + \frac{1}{2} - 2(k-1)\Psi_1 + 2\Psi_2, \quad (55)$$

where Ψ_2 is the second structure parameter that is defined in Equation (18). Finally, by combining the leading order expansions (50) and the moments (53) and (55), and by rescaling, as in Equation (31), we have the following explicit expressions for the two main

metrics of separability as functions of the multiplicity k and the structure parameters Ψ_1 and Ψ_2 :

$$\frac{p_c}{n} = \left(1 - \frac{1}{2k} - \frac{\Psi_1}{k}\right)^{-1}, \quad (56)$$

$$\frac{\Delta p}{\sqrt{n}} = 2k \left(k - \frac{1}{2} - \Psi_1\right)^{-\frac{3}{2}} \left(\frac{1}{4} + \Psi_1 - \Psi_1^2 + 2\Psi_2\right)^{\frac{1}{2}}. \quad (57)$$

For data that are structured as pairs of points, $k = 2$, Equation (56) gives the storage capacity of an ensemble of segments; this special result was first obtained, by means of replica calculations, in [44], and it was then rediscovered in other contexts in [8,45].

7.4. Dependence on the Structure Parameters and Scaling

The two structure parameters Ψ_1 and Ψ_2 , which control the two main metrics of linear separability, belong to k -dependent ranges:

$$\Psi_1 \in [0, k-1], \quad \Psi_2 \in [0, (k-1)(k-2)/2]. \quad (58)$$

The two quantities are not independent, since they are constructed from the same set of $k-1$ quantities $\psi_m \in [0, 1]$. When conditioned on a fixed value of Ψ_1 , Ψ_2 has a lower bound Ψ_2^- and an upper bound Ψ_2^+ that can be computed by considering the two following extreme cases. First, the supremum of Ψ_2 is realized in the maximum entropy case, where the value of Ψ_1 is uniformly distributed among the ψ_m . Second, the infimum of Ψ_2 corresponds to the minimum entropy case, where Ψ_1 is distributed on the fewest possible ψ_m 's. Explicitly,

$$\sup : \quad \{\psi_m\} = \left\{ \frac{\Psi_1}{k-1}, \dots, \frac{\Psi_1}{k-1} \right\}, \quad (59)$$

$$\inf : \quad \{\psi_m\} = \underbrace{\{1, \dots, 1\}}_{\lfloor \Psi_1 \rfloor}, \Psi_1 - \lfloor \Psi_1 \rfloor, 0, \dots, 0. \quad (60)$$

The definition of Ψ_2 , Equation (18), can be rewritten, as follows:

$$\Psi_2 = \frac{1}{2} \Psi_1^2 - \frac{1}{2} \sum_{m=2}^k \psi_m^2. \quad (61)$$

Substituting (59) and (60) into (61), we obtain

$$\Psi_2^+ = \sup \Psi_2 = \frac{\Psi_1^2}{2} \left(1 - \frac{1}{k-1}\right), \quad (62)$$

$$\Psi_2^- = \inf \Psi_2 = \Psi_1 \lfloor \Psi_1 \rfloor - \frac{1}{2} (\lfloor \Psi_1 \rfloor^2 + \lfloor \Psi_1 \rfloor). \quad (63)$$

Figure 3 shows the location of the transition, p_c , and the width of the region, Δp , as functions of Ψ_1 and Ψ_2 for a few values of k . Notice that the range of Δp at fixed k and Ψ_1 is itself bounded because of the limited range $[\Psi_2^-, \Psi_2^+]$ of Ψ_2 .

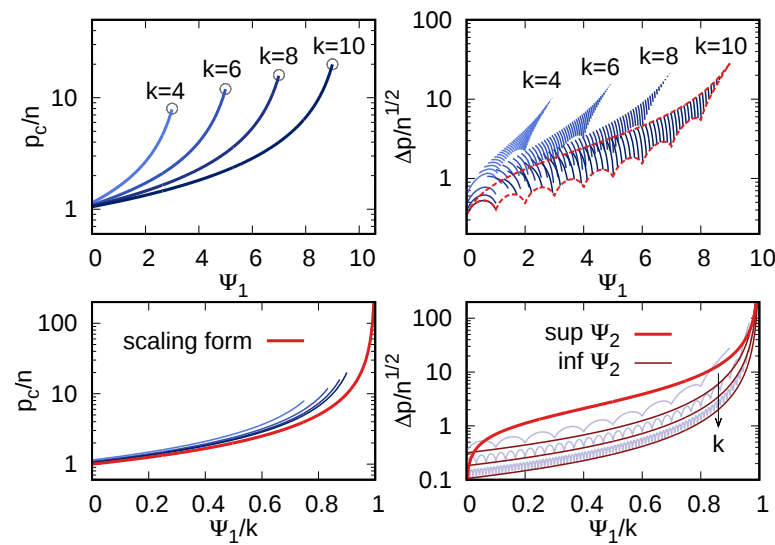


Figure 3. (Top left) The dependence of the rescaled location p_c/n of the transition region (y axis) on the structure parameter Ψ_1 (x axis), for a few values of the multiplicity k . Circles pinpoint the tips of the curves, which correspond to the unstructured case, where $\Psi_1 = k - 1$ (i.e., $\psi_m = 1$ for all m) and $p_c = 2kn$. (Top right) The rescaled width $\Delta p/\sqrt{n}$ of the transition region (y axis). Segments correspond to 50 fixed values of Ψ_2 , which were equally spaced in $[0, (k-1)(k-2)/2]$; their range in Ψ_1 (x axis) is obtained by inverting the relations (62) and (63). The dashed red lines are the upper and lower bounds of $\Delta p/\sqrt{n}$, obtained by substituting (62) and (63) into (57). (Bottom left) The large- k scaling form (red line) of p_c/n (y axis) as a function of the rescaled parameter Ψ_1/k (x axis); the blue lines are the same as in the top left panel. (Bottom right) The large- k behavior of the upper (thick red line) and lower (thin red and grey lines) bounds $\Delta p^\pm/\sqrt{n}$ (y axis) as functions of Ψ_1/k (x axis). Grey lines are the tight lower bounds as in the top right panel and thin red lines are the simpler bound Equation (66); different lower bounds correspond to $k = 10, 30, 90$.

There is an interesting observation to be made on a semi-quantitative level. At fixed k and n , p_c is an increasing function of Ψ_1 . The width Δp depends on both structure parameters, but, since the range of Ψ_2 at fixed Ψ_1 is so limited, one expects that, in practice, Δp will be approximately an increasing function of Ψ_1 . Therefore, Δp will be, in most cases, an increasing function of p_c . This is exactly the phenomenology that is observed in Figure 2, in both the synthetic data and MNIST.

The rescaled location of the transition p_c/n , Equation (56), does not depend on Ψ_2 , and it depends on Ψ_1 only through the rescaled value Ψ_1/k . For large k , it takes the scaling form

$$\frac{p_c}{n} \sim f_{p_c}\left(\frac{\Psi_1}{k}\right), \quad f_{p_c}(x) = \frac{1}{1-x}. \quad (64)$$

The width Δp , on the contrary, depends on both Ψ_1 and Ψ_2 . Because it is a monotonically increasing function of Ψ_2 , its upper bound Δp^+ and lower bound Δp^- at fixed Ψ_1 can be obtained by substituting (62) and (63) in Equation (57). Expressing Δp^+ again as a function of the rescaled parameter Ψ_1/k , and only keeping the leading term in $k \rightarrow \infty$, one obtains the scaling form

$$\frac{\Delta p^+}{\sqrt{n}} \sim f_{\Delta p^+}\left(\frac{\Psi_1}{k}\right), \quad f_{\Delta p^+}(x) = 2 \frac{\sqrt{x}}{1-x}. \quad (65)$$

Doing the same for Δp^- yields a complicated function, which is plotted in Figure 3. A simpler expression for the bound can be obtained by observing that $\Psi_2^- \geq (\Psi_1^2 - \Psi_1)/2$; using this more regular bound yields, at leading order in k ,

$$\frac{\Delta p^-}{\sqrt{n}} \sim k^{-\frac{1}{2}} f_{\Delta p^-}\left(\frac{\Psi_1}{k}\right), \quad f_{\Delta p^-}(x) = (1-x)^{-\frac{3}{2}}. \quad (66)$$

Figure 3 shows the large- k scaling behavior of p_c , Δp^+ , and Δp^- .

The two metrics are insensitive on most of the microscopic parameters of the theory, and they only depend on the two structure parameters, as shown analytically above. In addition, they display a large degree of robustness, even as functions of Ψ_1 and Ψ_2 : measuring p_c/n from the data fixes (up to corrections in k) the quantity Ψ_1/k , which, in turn, significantly narrows down the range of values that are attainable by Δp , the more so the smaller is k .

8. Discussion

The discussion above focused on the quantification of linear separability within a model that encodes simple relations between data points and their labels, in the form of constraints. Such a model has the advantage of being analytically tractable and allows the explicit expression of p_c and Δp in terms of model parameters. Moreover, the parameters appearing in the theory have direct interpretations as probabilities of geometric events, thus suggesting routes for further generalization.

In the face of its convenience for theoretical investigations, the definition of data structure used here does not aim at a realistic description of any specific data set. It must be interpreted as a phenomenological or effective parameterization of basic features of data structure that have a distinct effect on linear separability. The limited numerical experiments on MNIST data reported above are a proof of concept, showing a real data set with unexpectedly high linear separability, and they serve as a notable motivation for the investigation of data structure. The main goal of this article is the theoretical analysis; therefore, I postpone any comparison of theory and data. Moreover, MNIST is a relatively simple and clean data set. The numerical analysis signals the highly constrained nature of these data, where points that are close with respect to the Euclidean distance in $\mathbb{R}^n$ are more likely to have the same label. However, more complex data sets, such as ImageNET, are expected to be less constrained at the level of raw data, due to the higher variability within each category, and due to what are referred to as “nuisances”, i.e., elements that are present, but do not contribute to the classification. Yet, even in these cases, the aggregation of equally-labelled points emerges in the feature spaces towards the last layers of deep neural networks, which improves the efficacy of the linear readout downstream, as empirically observed [14,15].

An interesting, and perhaps unexpected, outcome of the theory concerns the universal properties of the probability of separation $c_{n,p}$. Here, I use the term “universality” in a much weaker sense than what is usually intended in statistical mechanics: I use it to denote (i) the qualitative robustness of the sigmoidal shape of the separability curve on the details of the model, and (ii) the quantitative insensitivity of the separability metrics on all but a few special combinations of parameters [46]. Importantly, the two metrics of data structure that are computed for the model, p_c and Δp , are the only two important parameters that fix $c_{n,p}$ in the thermodynamic limit, apart from the rescaling by k . The central limit theorem suggests this universality property. In fact, $\gamma_{n,q}$ is the probability distribution of the sum of $p - 1$ independent and identically distributed variables, as expressed by Equation (39). Therefore, $\gamma_{n,q}$ will converge to a Gaussian distribution with linearly increasing mean and variance. This indicates that μ_q and σ_q are the only two nonzero cumulants in the thermodynamic limit and, thus, q_c and Δq are the only two nontrivial metrics that are related to $\hat{c}_{n,q}$. This does not, by any means, imply that the model of data structure itself can be reduced to only two degrees of freedom. In fact, the phenomenology is richer if one considers the combinatorial quantity $C_{n,q}$ instead of the intensive one $\hat{c}_{n,q}$, see [32]; still, regarding the probability of separation, the relevant metrics are the location and width of the transition region.

Funding: The author acknowledges support from the University of Milan through the APC initiative.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The MNIST data set can be found here: yann.lecun.com/exdb/mnist/. (accessed on 20 May 2019)

Acknowledgments: I am grateful to Pietro Rotondo for countless stimulating discussions.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Boundary Conditions

The boundary conditions of the recurrence Equation (27) require some care. When a single ($q = 1$) multiplet is considered in dimension $n \geq k$, both its admissible dichotomies are linearly realizable. This is because all dichotomies of k points can be realized in $n \geq k$ dimensions, as I mentioned above. Therefore

$$\hat{c}_{n \geq k, 1} = 1. \quad (\text{A1})$$

The boundary conditions for $n < k$ are not simply the same as for $k = 1$. To see this, consider for instance what happens in $n = 1$ dimensions when dealing with a single ($q = 1$) multiplet of $k = 2$ points, ζ and $\bar{\zeta}$. Two problems arise: (i) if the two points lie on opposite sides of the origin, a linearly realized dichotomy ϕ will always assign them different signs, $\phi(\zeta) = -\phi(\bar{\zeta})$; (ii) there are not enough degrees of freedom to fix the overlap $\rho = \zeta \cdot \bar{\zeta}$ while keeping ζ and $\bar{\zeta}$ normalized.

These obstructions are problematic when trying to define the value of $\hat{c}_{1,1}$ for $k = 2$. This quantity appears in the right hand side of the recurrence Equation (25) when $n = 2$ and $q = 1$, where it is needed, alongside $\hat{c}_{2,1}$, to compute $\hat{c}_{2,2}$. Retracing the derivation for $k = 2$ shows that $\hat{c}_{1,1}$ in this context occurs when imposing a linear constraint in 2 dimensions, where it represents the fraction of admissible dichotomies of the doublet $\{\zeta, \bar{\zeta}\}$ that can be realized by a weight versor w satisfying $w \cdot \zeta' = 0$. In 2 dimensions, the orthogonality condition fixes w up to its sign. If this constrained w is such that

$$\text{sgn}(w \cdot \zeta) = \text{sgn}(w \cdot \bar{\zeta}) \quad (\text{A2})$$

then exactly 2 admissible dichotomies of $\{\zeta, \bar{\zeta}\}$ are realizable, otherwise the only realizable dichotomies are not admissible. Therefore $\hat{c}_{1,1}$ expresses the probability that Equation (A2) is satisfied; in the mean field approximation, this is $\psi_2(\rho)$. The foregoing argument actually applies for all $k \geq 1$. The probability that all k points in a multiplet lie in the same half-space with respect to the hyperplane realized by a random versor fixes the first non-trivial boundary condition $\hat{c}_{1,1}$.

For $k = 2$ this fixes everything. Let us now consider $k = 3$. In this case Equation (A1) omits $\hat{c}_{2,1}$. What should its value be? Again, going back to the argument in Section 6.3 is helpful. $\hat{c}_{2,1}$ appears in the recurrence when $n = 3$ and a linear constraint is imposed on w . This fixes w up to rotations around an axis, identified by a versor v . Now, whether the multiplet $\{\zeta^1, \zeta^2, \zeta^3\}$ allows 2 or 0 admissible dichotomies depends on whether there exists a vector w satisfying the constraint and such that $\text{sgn}(w \cdot \zeta^1) = \text{sgn}(w \cdot \zeta^2) = \text{sgn}(w \cdot \zeta^3)$. This happens if and only if the axis of rotation v lies outside the solid angle subtended by the three vectors $\zeta^1, \zeta^2, \zeta^3$. This characterization allows to compute $\hat{c}_{2,1}$ by elementary methods of solid geometry. One finds

$$\hat{c}_{2,1} = \frac{1}{2\pi} (1 - \omega_{123} - \omega_{231} - \omega_{312} - \pi) \quad [k = 3], \quad (\text{A3})$$

where

$$\omega_{abc} = \arccos \left(\frac{\rho_a - \rho_b \rho_c}{\sqrt{(1 - \rho_b^2)(1 - \rho_c^2)}} \right). \quad (\text{A4})$$

For larger values of k , the same reasoning allows to express the non trivial boundary conditions $\hat{c}_{n < k, 1}$ as geometric probabilities. Fortunately, the hassle of computing all these probabilities can be bypassed by using the boundary conditions (20), which are approximate

for $k > 1$, but still provide asymptotically correct results [28]. In fact, as is evident from the discussion in Section 7, if one takes the thermodynamic limit (30) the contribution of the $k - 1$ approximate values of $\hat{c}_{n,1}$ becomes negligible. Other ways of taking the thermodynamic limit (e.g., if k is extensive in n) may not enjoy this simplification, and may require a different analysis of the boundary conditions.

References

1. Yuan, G.X.; Ho, C.H.; Lin, C.J. Recent Advances of Large-Scale Linear Classification. *Proc. IEEE* **2012**, *100*, 2584–2603.
2. Elizondo, D. The linear separability problem: Some testing methods. *IEEE Trans. Neural Netw.* **2006**, *17*, 330–344, doi:10.1109/TNN.2005.860871.
3. Baldassi, C.; Della Vecchia, R.; Lucibello, C.; Zecchina, R. Clustering of solutions in the symmetric binary perceptron. *J. Stat. Mech. Theory Exp.* **2020**, *2020*, 073303.
4. Baldassi, C.; Malatesta, E.; Negri, M.; Zecchina, R. Wide flat minima and optimal generalization in classifying high-dimensional Gaussian mixtures. *J. Stat. Mech. Theory Exp.* **2020**, *2020*, 124012.
5. Cui, H.; Saglietti, L.; Zdeborová, L. Large deviations for the perceptron model and consequences for active learning. In Proceedings of the First Mathematical and Scientific Machine Learning Conference, Princeton, NJ, USA, 20–24 July 2020; pp. 390–430.
6. Aubin, B.; Perkins, W.; Zdeborová, L. Storage capacity in symmetric binary perceptrons. *J. Phys. Math. Theor.* **2019**, *52*, 294003, doi:10.1088/1751-8121/ab227a.
7. Gorban, A.N.; Tyukin, I.Y. Stochastic separation theorems. *Neural Netw.* **2017**, *94*, 255–259.
8. Chung, S.; Lee, D.D.; Sompolinsky, H. Linear readout of object manifolds. *Phys. Rev. E* **2016**, *93*, 060301.
9. Astrand, E.; Enel, P.; Ibos, G.; Dominey, P.F.; Baraduc, P.; Ben Hamed, S. Comparison of classifiers for decoding sensory and cognitive information from prefrontal neuronal populations. *PLoS ONE* **2014**, *9*, e86314, doi:10.1371/journal.pone.0086314.
10. Jacot, A.; Gabriel, F.; Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*; NIPS: Montreal, QC, Canada, 2018.
11. Lee, J.; Xiao, L.; Schoenholz, S.S.; Bahri, Y.; Novak, R.; Sohl-Dickstein, J.; Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *J. Stat. Mech. Theory Exp.* **2020**, *2020*, 124002, doi:10.1088/1742-5468/abc62b.
12. Saxe, A.M.; McClelland, J.L.; Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
13. Basu, M.; Ho, T. *Data Complexity in Pattern Recognition*; Springer: London, UK, 2006.
14. Cohen, U.; Chung, S.; Lee, D.D.; Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **2020**, *11*, 746, doi:10.1038/s41467-020-14578-5.
15. Ansuini, A.; Laio, A.; Macke, J.; Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems 32*; NIPS: Vancouver, BC, Canada, 2019.
16. Ingrosso, A. Optimal learning with excitatory and inhibitory synapses. *PLoS Comput. Biol.* **2021**, *16*, e1008536.
17. Zdeborová, L. Understanding deep learning is also a job for physicists. *Nat. Phys.* **2020**, *16*, 602–604, doi:10.1038/s41567-020-0929-2.
18. Bahri, Y.; Kadmon, J.; Pennington, J.; Schoenholz, S.S.; Sohl-Dickstein, J.; Ganguli, S. Statistical Mechanics of Deep Learning. *Annu. Rev. Condens. Matter Phys.* **2020**, *11*, 501–528, doi:10.1146/annurev-conmatphys-031119-050745.
19. Goldt, S.; Mézard, M.; Krzakala, F.; Zdeborová, L. Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Phys. Rev. X* **2020**, *10*, 041044, doi:10.1103/PhysRevX.10.041044.
20. Erba, V.; Ariosto, S.; Gherardi, M.; Rotondo, P. Random geometric graphs in high dimension. *Phys. Rev. E* **2020**, *102*, 012306, doi:10.1103/PhysRevE.102.012306.
21. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
22. Poole, B.; Lahiri, S.; Raghu, M.; Sohl-Dickstein, J.; Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 26*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds; Curran Associates, Inc.: New York, NY, USA, 2016; Volume 29, pp. 3360–3368.
23. Mazzolini, A.; Gherardi, M.; Caselle, M.; Cosentino Lagomarsino, M.; Osella, M. Statistics of Shared Components in Complex Component Systems. *Phys. Rev. X* **2018**, *8*, 021023, doi:10.1103/PhysRevX.8.021023.
24. Mazzolini, A.; Grilli, J.; De Lazzari, E.; Osella, M.; Lagomarsino, M.C.; Gherardi, M. Zipf and Heaps laws from dependency structures in component systems. *Phys. Rev. E* **2018**, *98*, 012315, doi:10.1103/PhysRevE.98.012315.
25. Gherardi, M.; Rotondo, P. Measuring logic complexity can guide pattern discovery in empirical systems. *Complexity* **2016**, *21*, 397–408.
26. Mézard, M. Mean-field message-passing equations in the Hopfield model and its generalizations. *Phys. Rev. E* **2017**, *95*, 022117, doi:10.1103/PhysRevE.95.022117.
27. Chung, S.; Lee, D.D.; Sompolinsky, H. Classification and Geometry of General Perceptual Manifolds. *Phys. Rev. X* **2018**, *8*, 031003, doi:10.1103/PhysRevX.8.031003.
28. Rotondo, P.; Cosentino Lagomarsino, M.; Gherardi, M. Counting the learnable functions of geometrically structured data. *Phys. Rev. Res.* **2020**, *2*, 023169, doi:10.1103/PhysRevResearch.2.023169.

29. Erba, V.; Gherardi, M.; Rotondo, P. Intrinsic dimension estimation for locally undersampled data. *Sci. Rep.* **2019**, *9*, 17133, doi:10.1038/s41598-019-53549-9.
30. Facco, E.; d'Errico, M.; Rodriguez, A.; Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **2017**, *7*, 12140, doi:10.1038/s41598-017-11873-y.
31. Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326, doi:10.1126/science.290.5500.2323.
32. Pastore, M.; Rotondo, P.; Erba, V.; Gherardi, M. Statistical learning theory of structured data. *Phys. Rev. E* **2020**, *102*, 032119, doi:10.1103/PhysRevE.102.032119.
33. Rotondo, P.; Pastore, M.; Gherardi, M. Beyond the Storage Capacity: Data-Driven Satisfiability Transition. *Phys. Rev. Lett.* **2020**, *125*, 120601, doi:10.1103/PhysRevLett.125.120601.
34. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278.
35. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999.
36. Schläfli, L. *Gesammelte Mathematische Abhandlungen*; Springer: Basel, Switzerland, 1950.
37. Cameron, S. *An Estimate of the Complexity Requisite in a Universal Decision Network*; WADD Technical Report; Bionics Symposium: Dayton, OH, USA, 1960.
38. Cover, T.M. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Trans. Electron. Comput.* **1965**, *EC-14*, 326–334, doi:10.1109/PGEC.1965.264137.
39. Baldi, P. A computational theory of surprise. In *Information, Coding and Mathematics: Proceedings of Workshop Honoring Prof. Bob McEliece on His 60th Birthday*; Blaum, M., Farrell, P.G., van Tilborg, H.C.A., Eds.; The Springer International Series in Engineering and Computer Science; Springer: Boston, MA, USA, 2002; pp. 1–25.
40. McCoy, B.M. *Advanced Statistical Mechanics*; Oxford University Press: Oxford, UK, 2010.
41. Caracciolo, S.; Di Gioacchino, A.; Gherardi, M.; Malatesta, E.M. Solution for a bipartite Euclidean traveling-salesman problem in one dimension. *Phys. Rev. E* **2018**, *97*, 052109, doi:10.1103/PhysRevE.97.052109.
42. Dell'Aquila, G.; Ferrante, M.I.; Gherardi, M.; Cosentino Lagomarsino, M.; Ribera d'Alcalà, M.; Iudicone, D.; Amato, A. Nutrient consumption and chain tuning in diatoms exposed to storm-like turbulence. *Sci. Rep.* **2017**, *7*, 1828, doi:10.1038/s41598-017-02084-6.
43. Gherardi, M.; Amato, A.; Bouly, J.P.; Cheminant, S.; Ferrante, M.I.; d'Alcalá, M.R.; Iudicone, D.; Falciatore, A.; Cosentino Lagomarsino, M. Regulation of chain length in two diatoms as a growth-fragmentation process. *Phys. Rev. E* **2016**, *94*, 022418, doi:10.1103/PhysRevE.94.022418.
44. Lopez, B.; Schroder, M.; Oppen, M. Storage of correlated patterns in a perceptron. *J. Phys. Math. Gen.* **1995**, *28*, L447.
45. Borra, F.; Lagomarsino, M.C.; Rotondo, P.; Gherardi, M. Generalization from correlated sets of patterns in the perceptron. *J. Phys. Math. Theor.* **2019**, *52*, 384004, doi:10.1088/1751-8121/ab3709.
46. Machta, B.B.; Chachra, R.; Transtrum, M.K.; Sethna, J.P. Parameter Space Compression Underlies Emergent Theories and Predictive Models. *Science* **2013**, *342*, 604, doi:10.1126/science.1238723.