Generalization from correlated sets of patterns in the perceptron

Francesco Borra[1], Marco Cosentino Lagomarsino[2,3], Pietro Rotondo[4,5], and Marco Gherardi[2,6]

[1]Università degli studi di Roma La Sapienza, Italy
[2]Università degli studi di Milano, via Celoria 16, Milano, Italy
[3]IFOM Foundation FIRC Institute of Molecular Oncology, Milan Italy
[4]School of Physics and Astronomy, University of Nottingham, Nottingham, NG7 2RD, UK
[5]Centre for the Mathematics and Theoretical Physics of Quantum Non-equilibrium Systems, University of Nottingham, Nottingham NG7 2RD, UK
[6]INFN Sezione di Milano

July 3, 2019

Abstract

Generalization is a central aspect of learning theory. Here, we propose a framework that explores an auxiliary *task-dependent* notion of generalization, and attempts to quantitatively answer the following question: given two sets of patterns with a given degree of dissimilarity, how easily will a network be able to "unify" their interpretation? This is quantified by the volume of the configurations of synaptic weights that classify the two sets in a similar manner. To show the applicability of our idea in a concrete setting, we compute this quantity for the perceptron, a simple binary classifier, using the classical statistical physics approach in the replica-symmetric ansatz. In this case, we show how an analytical expression measures the "distance-based capacity", the maximum load of patterns sustainable by the network, at fixed dissimilarity between patterns and fixed allowed number of errors. This curve indicates that generalization is possible at any distance, but with decreasing capacity. We propose that a distance-based definition of generalization may be useful in numerical experiments with real-world neural networks, and to explore computationally sub-dominant sets of synaptic solutions.

1 Introduction

Generalization is an essential feature of cognition. It constructs broad, universal statements or general concepts from a few empirical observations. Our ability to generalize comprises a wide and not well characterized set of tasks and abilities, including the tendency to respond in the same way to different, but similar "on some level", stimuli. An important feature of generalization is to assume or recognize the existence of common features shared by sets of elements. Hence, classifying common relations and differences among different observed patterns is crucial [1].

The problem of breaking down the process of generalization into its elementary components naturally emerges in the field of artificial neural networks. In this context, achieving a deeper understanding of generalization could improve the current design principles and clarify the reasons why current architectures generalize well. A generally held belief is that the efficiency of deep neural networks in identifying patterns can be understood in terms of feature extraction. Despite its robustness, this view is being challenged by an increasing number of experimental results. For example, a recent study showed that this paradigm is misleading when trying to understand how deep neural networks generalize, and proposes to consider alternative approaches [2]; one such approach already emerged in a statistical mechanics setting [3, 4]. Another challenging observation is a recently discovered fragility of neural networks to so-called "adversarial attacks" [5], whereby any given pattern can be maliciously modified into a very similar pattern that gets misclassified.

Learning efficiently from examples requires a way to gauge the size and quality of the training and test sets. The classic methods to address these issues (e.g., bias-variance decomposition, Vapnik-Chervonenkis dimension, model complexity) are found within so-called "statistical learning theory" [6]. In this framework, the goal of learning is the approximation of a function f by an element h of a given hypothesis space $\mathcal{H}$. The trained machine h is chosen by a learning algorithm, starting from a training set, i.e. a set of pairs $\xi_\mu, f(\xi_\mu)$, where ξ_μ are chosen from a space Ω with unknown probability distribution. Statistical learning theory establishes general relations between the complexity or expressivity of a given function class $\mathcal{H}$ and its generalization properties, defined as the ability of its elements to reproduce the output of f beyond the training set [7]. Despite its power, this framework has limitations concerning its applicability to neural networks [8]. For instance, it was observed recently that over-parameterized neural networks often generalize better than smaller, less complex, ones. This finding is in conflict with the predictions of statistical learning and classical computational complexity theory (as well as with naive intuition) [9].

Another important drawback of statistical learning theory is that it considers generalization in a worst-case scenario, where the capabilities of the network are tested against possibly malicious counterexamples [10]. A statistical physics approach overcomes this problem by considering the so-called "teacher-student" scenario [4, 11, 12]. This framework usually assumes that the generalization

ability of a trained network (the student) is measured on a set of tests produced by a teacher with identical architecture

(although it is possible to address mismatched architectures as well).

An important limitation of the teacher-student setting is that the student can learn from irrelevant examples if the teacher is used as a surrogate for a classification rule not involving the whole space of inputs. Consider, for instance, a teacher trained to discern handwritten digits from a training set $\Omega_{\rm t}$ (e.g. from the popular MNIST database). The training set is a subset of a much larger set Ω of all black-and-white images. The teacher will associate well-defined output values to elements of $\Omega_{\rm t}$ but also to any other element of Ω, for instance to those in a set $\Omega_{\rm n}$, disjoint from $\Omega_{\rm t}$, containing random noise. Now the student can in principle learn to reproduce the teacher's response on $\Omega_{\rm t}$ by learning to mimic the teacher on the (meaningless) set $\Omega_{\rm n}$. In other words, the student can learn to classify a dataset without ever seeing a single example from it.

This problem is connected to the more general fact that the space Ω, in practical applications, has structure, first and foremost via notions of distance or similarity between its elements.

How such structure affects neural computation is a question that is becoming more and more prominent, especially in the statistical physics literature, owing to the fact that (i) real datasets usually have non-trivial features, and (ii) structure affects the functioning of machine learning algorithms [13, 14, 15, 16, 17]. Moreover, data are not noise-free nor devoid of corruption: understanding the connections between the geometry of Ω and the performances of learning algorithms would allow to tackle more quantitatively the issues of learning noisy data.

Building on these premises, we propose here an auxiliary way to approach generalization.

We will use the term "generalization" in a loose way, to clarify our intents, but with no straightforward relation to the traditional definitions.

The main object of interest is the number of different synaptic configurations that allow the network to classify correctly two sets of patterns. This quantity is an extension of the classic "Gardner volume" [18, 19] to the case where the training set consists of two sets of patterns with a prescribed degree of correlation. The rationale is to take explicitly into account the degree of similarity between "new" and "old" information seen by a network. Figure 1 is a stylized sketch of our framework. Here we use a simple definition of distance between binary data sets, based on their overlaps, but the framework applies to any other definition of distance. The load at which the Gardner volume becomes zero defines a capacity, which in our case depends on the distance.

This approach aims at a data-dependent, or task-dependent, concept of generalization; at the same time it is in some sense "rule agnostic" (in the same spirit of statistical learning theory) as it does not implement an explicit generalization test (such as the teacher-student).

To root our ideas in a concrete case, we carry out these calculations on a very simple neural network, the perceptron, using classic results from statistical mechanics [18, 20].

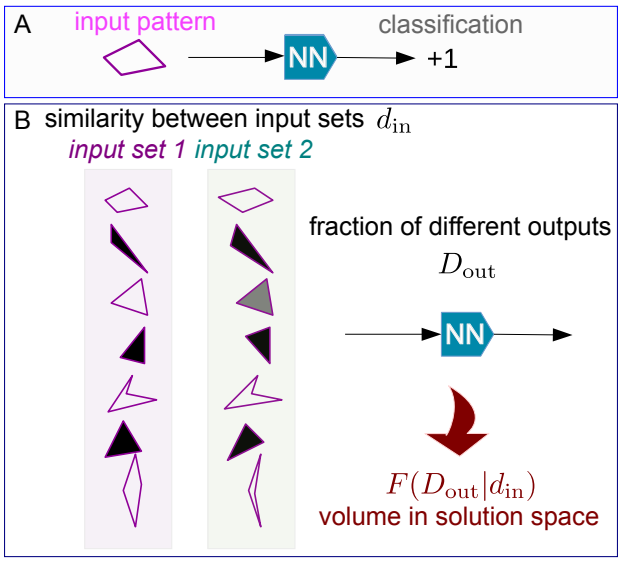

Figure 1: Generalization is related to the ability of a network to classify similar inputs alike. We address this problem by studying the number of network configurations (its logarithm is proportional to F) which classify two input sets, whose similarity is quantified by $d_{\rm in}$, with $D_{\rm out}$ output mismatches.

2 Definition of the approach and results

2.1 Measuring the similarity between classification tasks

This section defines two complementary notions of dissimilarity between classification tasks, one based on a bit-wise comparison of the training sets, and the other based on the free-energy landscape of the corresponding supervised learning problem.

A classifier neural network is described by an output function σ that associates an output $\sigma(\boldsymbol{\xi}_\mu)$ to any input vector $\boldsymbol{\xi}_\mu \in \mathbb{R}^N$. We consider binary input vectors, whose elements are ± 1 bits and we define a classification task as a training set $\xi = \{\boldsymbol{\xi}_\mu\}_{\mu=1,\dots,p}$ with associated labels $\sigma = \{\sigma_\mu\}_{\mu=1,\dots,p}$. Therefore, a task is a set of input-output pairs $(\xi, \sigma) = \{(\boldsymbol{\xi}_\mu, \sigma_\mu)\}_{\mu=1,\dots,p}$. The network can solve the task if $\forall \mu = 1, \dots, p$, $\sigma_\mu = \sigma(\boldsymbol{\xi}_\mu)$.

Let us fix two classification tasks, say (ξ, σ) and $(\bar{\xi}, \bar{\sigma})$. We focus on inputs ($\xi$ and $\bar{\xi}$) and outputs (σ and $\bar{\sigma}$) separately, thus defining two bit-wise distances. The canonical distance between two patterns $\boldsymbol{\xi}_\mu$ and $\boldsymbol{\xi}_\nu$ is the normalized Hamming distance, defined as

$$d_{\rm H}(\boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu) = \frac{1}{2} - \frac{1}{2N} \boldsymbol{\xi}_\mu \cdot \boldsymbol{\xi}_\nu. \tag{1}$$

The quantity $\boldsymbol{\xi}_\mu \cdot \boldsymbol{\xi}_\nu / N = \sum_{i=1}^N \xi_\mu^i \xi_\nu^i / N =: q(\boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu)$ is the overlap between

$\boldsymbol{\xi}_\mu$ and $\boldsymbol{\xi}_\nu$. We need to extend this definition to the full sets of inputs $\xi = \{\boldsymbol{\xi}_\mu\}_{\mu=1,\dots,p}$ and $\bar{\xi} = \{\bar{\boldsymbol{\xi}}_{\bar{\mu}}\}_{\bar{\mu}=1,\dots,p}$. Intuitively, ξ and $\bar{\xi}$ are identical if, for any $\boldsymbol{\xi}_\mu \in \xi$ there exists some $\bar{\boldsymbol{\xi}}_{\bar{\mu}} \in \bar{\xi}$ such that $\boldsymbol{\xi}_\mu = \bar{\boldsymbol{\xi}}_{\bar{\mu}}$. Following this line of reasoning, one can define a distance between the sets ξ and $\bar{\xi}$ as

$$d_{\text{in}}(\xi, \bar{\xi}) = \min_{\pi \in S_p} \sum_\mu d(\boldsymbol{\xi}_\mu, \bar{\boldsymbol{\xi}}_{\pi(\mu)}), \tag{2}$$

where S_p is the symmetric group of order p. The minimum over S_p selects the closest matching between each pattern in ξ and each one in $\bar{\xi}$, thus restoring the symmetry under permutations, but it complicates the analytical computations. Therefore, we adopt a simpler setup, still inspired by (2), where the elements of the input sets are at fixed distance d pairwise, meaning that $d_{\text{in}}(\xi_\mu, \bar{\xi}_\mu) = d_{\text{in}}$ for every μ. We consider the ensemble of p independent pairs of inputs at Hamming distance d_{in}, defined by the joint probability

$$P_{d_{\text{in}}}(\xi, \bar{\xi}) := P(\xi, \bar{\xi} | d_{\text{in}}(\xi, \bar{\xi}) = d_{\text{in}}) \tag{3}$$

$$= \frac{1}{P(d_{\text{in}})^p} \prod_{\mu=1}^{p} P(\boldsymbol{\xi}_\mu) P(\bar{\boldsymbol{\xi}}_\mu)\, \delta(d_{\text{H}}(\boldsymbol{\xi}_\mu, \bar{\boldsymbol{\xi}}_\mu) - d_{\text{in}}) \tag{4}$$

where $P(d_{\text{in}}) := P(d_{\text{H}}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = d_{\text{in}})$ is the so-called Bayesian normalization, i.e., the probability that a pair of two random vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ has Hamming distance d_{in}.

Notice that this ensemble, fixing the distance between ξ and $\bar{\xi}$, can be interpreted as a model of noisy input data, at least for small values of d_{in}. In fact, if one considers a fixed ξ, then the probability $P_{d_{\text{in}}}(\xi, \bar{\xi})$ is concentrated on all inputs $\bar{\xi}$ which are similar (but not identical) to ξ.

Having fixed the matching between inputs by this definition, it is then natural to define the distance between outputs $\sigma = \{\sigma_\mu\}_{\mu=1,\dots,p}$ and $\bar{\sigma} = \{\bar{\sigma}_{\bar{\mu}}\}_{\bar{\mu}=1,\dots,p}$ as

$$d_{\text{out}}(\sigma, \bar{\sigma}) = \frac{1}{p} \sum_{\mu=1}^{p} (1 - \delta_{\sigma_\mu, \bar{\sigma}_\mu}). \tag{5}$$

Analogously to what mentioned above, an ensemble fixing $d_{\text{out}}(\sigma, \bar{\sigma})$ can be interpreted as modeling noise in the input labels (e.g., mislabeling).

We now focus on the similarity between two tasks in terms of synaptic representation. We reason that, from the point of view of the network, two tasks are similar if they are easily solved by the same synaptic configuration, i.e., if they share a significant fraction of common solutions. Here, by solutions we mean the solutions to the problem of finding a state of the network, specified by a synaptic weight structure W, such that the task is solved correctly. Two equivalent tasks will share all solutions and should have zero synaptic distance. Conversely, two tasks are incompatible if they have no common solution, or, equivalently, if there exists no solution to the task of learning the union of the two tasks. In this section, we do not specify the nature of W any further, since

our approach can be applied to more general multi-layered architectures. In the following sections we apply our ideas to the case of the perceptron only.

A definition of synaptic distance consistent with these intuitive requirements can be formalized as follows. Let us consider a neural network with cost function $H_{\xi,\sigma}(W)$, equal, for instance, to the number of errors that the network makes in performing the task (ξ, σ) when its synaptic weights are W. Let the cost function be additive under union of the training sets: given two sets of input-output pairs (ξ, σ) and $(\bar{\xi}, \bar{\sigma})$ of arbitrary sizes p and $\bar{p}$,

$$H_{(\xi,\sigma)\cup(\bar{\xi},\bar{\sigma})}(W) = H_{\xi,\sigma}(W) + H_{\bar{\xi},\bar{\sigma}}(W), \tag{6}$$

where $(\xi, \sigma) \cup (\bar{\xi}, \bar{\sigma})$ denotes the labelled training set with inputs $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p, \bar{\boldsymbol{\xi}}_1, \ldots, \bar{\boldsymbol{\xi}}_{\bar{p}}\}$ and outputs $\{\sigma_1, \ldots, \sigma_p, \bar{\sigma}_1, \ldots, \bar{\sigma}_{\bar{p}}\}$. The canonical partition function at inverse temperature β of a system with Hamiltonian $H_{\xi,\sigma}$ is

$$Z(\beta, (\xi, \sigma)) = \int dW \; e^{-\beta H_{\xi,\sigma}(W)}, \tag{7}$$

where dW is a normalized measure on the synaptic weights. In the zero-temperature limit $\beta \to \infty$, the system occupies the lowest-energy state and Z becomes the degeneracy of the ground state, or the fraction of exact solutions to the task (ξ, σ). The free energy is, in this context, referred to as the Gardner volume:

$$F(\beta, (\xi, \sigma)) = -\frac{1}{\beta N} \ln \, Z(\beta, (\xi, \sigma)) \tag{8}$$

Our definition of network distance is then

$$\Omega_{\mathrm{n}}((\xi, \sigma), (\bar{\xi}, \bar{\sigma}), \beta) := \frac{1}{\beta N} \ln \frac{\sqrt{Z(\beta, (\xi, \sigma)) Z(\beta, (\bar{\xi}, \bar{\sigma}))}}{Z(\beta, (\xi, \sigma) \cup (\bar{\xi}, \bar{\sigma}))}, \tag{9}$$

(where the subscript n stands for "network").

In the zero-temperature limit $\beta \to \infty$ this quantity counts the number of exact common solutions to the two tasks, normalized by the number of solutions to the two tasks separately. It must be remarked that, in the SAT phase, the beta prefactor in (9) should be removed when taking the zero temperature limit, in order to recover a finite quantity (the entropy) as the free energy would be zero. In fact, Eq. (9) can be rewritten in terms of the free energy as

$$\Omega_{\mathrm{n}}((\xi, \sigma), (\bar{\xi}, \bar{\sigma}), \beta) = -\frac{F(\xi, \sigma) + F(\bar{\xi}, \bar{\sigma})}{2} + F((\xi, \sigma) \cup (\bar{\xi}, \bar{\sigma})) \tag{10}$$

The synaptic distance is zero for two identical tasks and diverges for incompatible tasks: $\Omega_{\mathrm{n}}((\xi, \sigma), (\xi, \sigma), \infty) = 0$ and $\Omega_{\mathrm{n}}((\xi, \sigma), (\xi, -\sigma), \infty) = \infty$.

In this setting, the ability of the network to generalize can be studied by comparing the two types of distances defined above. In particular, we will consider the typical value of Ω in the ensemble where $d_{\mathrm{in}} = d_{\mathrm{in}}(\xi, \bar{\xi})$ and $D_{\mathrm{out}} = d_{\mathrm{out}}(\sigma, \bar{\sigma})$ are fixed:

$$\Omega(D_{\mathrm{out}}, d_{\mathrm{in}}) = \left\langle \Omega_{\mathrm{n}}((\xi, \sigma), (\bar{\xi}, \bar{\sigma}), \beta) \right\rangle_{d_{\mathrm{in}}(\xi,\bar{\xi})=d_{\mathrm{in}};\;\; d_{\mathrm{out}}(\sigma,\bar{\sigma})=D_{\mathrm{out}}}. \tag{11}$$

A critical line is identified by the point in which $\Omega = \infty$ for fixed $(D_{\text{out}}, d_{\text{in}})$ for a given size of the input sets. Equivalently, for fixed d_{in} and size p, $\Omega = \infty$ indicates the theshold values of D_{out}, i.e. the range of output-similarity that the network can typically attribute to the input sets.

2.2 Generalization properties of the perceptron

We now set out to specify the abstract notion of memory-based distance introduced in the previous section (Eq. (11)) in order to use it for an explicit calculation. We call $\xi = \{\boldsymbol{\xi}_\mu\}_{\mu=1,\dots,p}$ a set of $p = \alpha N$ input vectors with components $\xi_\mu^i = \pm 1 \quad \forall i = 1, ..., N$. The perceptron is a network which yields a binary output $\sigma_W(\xi_\mu) = \text{sign}(\boldsymbol{W} \cdot \boldsymbol{\xi}_\mu / N - K) = \pm 1$ for any input, given a certain synaptic configuration $\boldsymbol{W} \in \mathbb{R}^N$. In the spherical model $\sum_i W_i^2 = N$, in the discrete model $W_i = \pm 1$.

In the case of batch learning, for any given training set (ξ, σ), the energy conventionally associated to this model is the error-counting cost function

$$H_{\xi,\sigma}(\boldsymbol{W}) = \sum_{\mu=1}^{p} \Theta(-\sigma_\mu \ \ \boldsymbol{\xi}_\mu \cdot \boldsymbol{W}). \tag{12}$$

This definition allows to compute the Gardner volume, i.e. the fraction of synaptic configurations $\boldsymbol{W}$ that solve a certain task, as defined in (8).

As for the probabily of synaptic configurations $dP(\boldsymbol{W})$, the following maximally entropic probability distributions are conventionally used, for the spherical and discrete case respectively

$$\begin{aligned} dP(\boldsymbol{W}) &= \frac{d\boldsymbol{W}\ \delta\left(\sum_{i=1}^N W_i^2 - N\right)}{\int_{\mathbb{R}^N} d\boldsymbol{W}\ \delta\left(\sum_{i=1}^N W_i^2 - N\right)} \quad \text{(spherical)} \\ dP(\boldsymbol{W}) &= \frac{1}{2^N} \quad \text{(discrete)}. \end{aligned} \tag{13}$$

A different cost function was proposed recently, in order to study a peculiar clustering property of the solutions in the perceptron with discrete weigths [21, 22, 23].

In order to study the typical behaviour of this network, the Gardner volume should be averaged on the input-output pair statistics $P(\xi, \sigma)$ (quenched average). A second-order phase transition is witnessed by the average value of the cost function. The critical capacity $\alpha_c(\beta)$ is defined as

$$\alpha_c(\beta) = \inf_\alpha \{\alpha : \ \langle H \rangle_{(\xi,\sigma)} > 0\} \tag{14}$$

and identifies a critical line in the (α, β) plane. Physically, $\alpha_c(\beta)$ is the maximum number of patterns per neuron that can be learned in the typical case (without committing an extensive number of mistakes). The value of α_c clearly depends

on the statistics of the input-ouput pairs. The simplest statistics is obtained by choosing both inputs and outputs randomly and independently

$$P(\xi,\sigma)=\prod_{\mu=1}^{p}[a\,\delta_{\sigma_\mu,-1}+(1-a)\,\delta_{\sigma_\mu,1}]\prod_{j=1}^{N}[b\,\delta_{\xi_\mu^j,-1}+(1-b)\,\delta_{\xi_\mu^j,1}]\;. \tag{15}$$

In the unbiased case, $a=b=1/2$ and $\beta=\infty$, $\alpha_c=2$ for the spherical perceptron, while $\alpha_c\approx 0.833$ in the discrete case. It is important to point out that, in the latter case, the replica-symmetric ansatz yields a quantitatively incorrect result and one-step replica symmetry breaking is needed [24].

In the case of a single set of independent and spatially correlated inputs, two classic studies [25, 26], derive the capacity from the (intra) correlation matrix $C_{ij}=\langle\xi_\mu^i\xi_\mu^j\rangle$. More recent studies [27, 28] have focused on the capacity in the case of a prescribed correlation between different patterns, as given by the overlap matrix $C_{\mu\nu}=\boldsymbol{\xi}_\mu\cdot\boldsymbol{\xi}_\nu/N$.

The teacher-student setting for generalization assumes the following specific choice of the output statistics. Instead of drawing inputs and outputs independently, there is an input (example) statistics $P(\xi)$ and a teacher machine, described by an output function σ_T. For each input, the "correct" output is chosen by the teacher. Specifically, the input-output statistics is given by

$$P(\xi,\sigma)=P(\xi)\prod_{\mu=1}^{p}\delta(\sigma_\mu-\sigma_T(\xi_\mu)) \tag{16}$$

in a noise-free scenario. In this scenario, the cost function (called "learning function"), is algorithm-specific, and its average value $\langle H\rangle$, as a function of the number of examples $\alpha=p/N$, is called "learning curve" $\epsilon(\alpha)$. The learning curve quantifies generalization. The generalization ability can be defined by averaging on all possible teachers, if needed. For a perceptron, the best possible performance is achieved with the Bayesian algorithm which, however, can only be performed by a perceptron-based committee machine [29, 30, 20].

2.2.1 Distance-based Gardner volume

Turning to our approach to define generalization, we choose both inputs ξ and outputs σ to be random and unbiased. Since we have two sets (ξ,σ) and $(\bar\xi,\bar\sigma)$, the standard gauge freedom of the problem allows to fix the outputs of the first set σ_μ to $+1$. With this premise, we write the non-trivial part of Eq. (10), for given $d_{\rm out}(\sigma,\bar\sigma)=D_{\rm out}$, as

$$F(D_{\rm out})=\frac{1}{N}\ln\int dP(\boldsymbol{W}) \tag{17}$$

$$\sum_{\{\epsilon_\mu=\pm1\}}\delta\Big(\sum_{\mu=1}^{p}\epsilon_\mu-p(1-2D_{\rm out})\Big)\prod_{\mu=1}^{p}\Theta(\boldsymbol{W}\cdot\boldsymbol{\xi}_\mu)\,\Theta(\epsilon_\mu\boldsymbol{W}\cdot\bar{\boldsymbol{\xi}}_{\bar\mu}). \tag{18}$$

Note that this is a "zero-temperature" definition, and that

$$\sum_{\{\epsilon_\mu=\pm 1\}} \delta\Big(\sum_{\mu=1}^{p}\epsilon_\mu - p(1-2D_{\rm out})\Big)\ \prod_{\mu=1}^{p}\Theta(\boldsymbol{W}\cdot\boldsymbol{\xi}_\mu)\,\Theta(\epsilon_\mu \boldsymbol{W}\cdot\bar{\boldsymbol{\xi}}_{\bar{\mu}})$$

equals one if exactly $D_{\rm out}p$ output pairs are discordant, regardless of which pairs, and zero otherwise. This choice of summing over ϵ_μ inside the logarithm makes Eq. (17) slightly different from Eq. (10). Nonetheless, the two quantities are equal up to a combinatorial prefactor which is irrelevant for the final result. This also justifies the conditional notation $(D_{\rm out}|d_{\rm in})$ once the average over input sets of given distance is taken.

The full expression for (11) is therefore

$$\Omega(D_{\rm out}|d_{\rm in}) = \overline{F_\alpha} - \overline{F_\alpha(D_{\rm out}|d_{\rm in})} \tag{19}$$

where $\overline{F_\alpha} = \overline{F_\alpha(0|0)}$ is the conventional Gardner volume for $N\alpha$ inputs, and

$$\overline{F_\alpha(D_{\rm out}|d_{\rm in})} = \int P_{d_{\rm in}}(\xi,\bar{\xi})F_\alpha(\xi,\bar{\xi},D_{\rm out})\,. \tag{20}$$

Both these observables can be evaluated within the replica formalism, as will be outlined in the next section (a detailed derivation will be given in Appendix A and B).

Finally, we introduce the distance-based capacity $\alpha_c(D_{\rm out}|d_{\rm in})$, which is to $\overline{F_\alpha(D_{\rm out}|d_{\rm in})}$ what α_c (Eq. (14)) is to the Gardner volume (8). Since in our "zero temperature" framework we have not explicitly introduced a cost function, it is convenient to define $\alpha_c(D_{\rm out}|d_{\rm in})$ as

$$\alpha_c(D_{\rm out}|d_{\rm in}) = \inf_\alpha\{\alpha : \overline{F_\alpha(D_{\rm out}|d_{\rm in})} > -\infty\}. \tag{21}$$

Physically, $\alpha_c(D_{\rm out}|d_{\rm in})N$ is the maximum size of two sets, with distance $d_{\rm in}$, that can be learned simultaneously with $D_{\rm out}$ $\alpha_c(D_{\rm out}|d_{\rm in})N$ concordant output pairs.

Now, suppose the perceptron has learned a set ξ of size αN. If a new set $\bar{\xi}$, with $d_{\rm in}(\xi,\bar{\xi}) = d_{\rm in}$ is presented, then there will be a fraction $D_{\rm out}$ of discordant outputs and in this case $\alpha_c(D_{\rm out}|d_{\rm in})$ shows the range $D_{\rm out}$ can assume, given $d_{\rm in}$. It must be remarked that this boundary does not show whether any $D_{\rm out}$ has a finite probability. A related quantity is the conditional probability $P(D_{\rm out}|d_{\rm in})$, which may be approximated by

$$P(D_{\rm out}|d_{\rm in}) \approx \frac{\exp(NF(D_{\rm out}|d_{\rm in}))}{\int dD'\ \exp(NF(D'|d_{\rm in}))} \tag{22}$$

$$= \exp[NF(D_{\rm out}|d_{\rm in}) - N\max_{D'}F(D'|d_{\rm in})] \to \begin{cases}1 & \text{finite chance}\\ 0 & \text{no chance}\end{cases} \tag{23}$$

However, this idea requires some non-trivial computations and careful geometrical considerations, and we do not pursue it here.

2.2.2 Analytical expression for the distance-based capacity

We are able to derive an analytical expression for the critical capacity at fixed pattern distance, which we discuss in the following. The quantity $\overline{F_\alpha(D_{\text{out}}|d_{\text{in}})}$ in Eq. (20) is, formally, an average free energy and can be computed with the replica formalism [18, 20]. The replica method is based on the identity

$$\overline{F_\alpha(D_{\text{out}}|d_{\text{in}})} = \int P(\xi) \ln Z = \lim_{m\to 0} \frac{1}{m} \ln \int P(\xi)\ Z^m \tag{24}$$

Following standard procedure, we introduce the parameter Q_{ab} by

$$1 = \int \prod_{a<b} dQ_{ab}\ \delta(Q_{ab} - \boldsymbol{W}_a \cdot \boldsymbol{W}_b/N). \tag{25}$$

Hence

$$\int P(\xi)\ Z^m = \int P(\xi)\ \int \prod_{a=1}^{m} d\boldsymbol{W}_a\ e^{-\beta H_\xi(\boldsymbol{W}_a)} =: \int \prod_{a<b} dQ_{ab}\ e^{NA[\{Q_{ab}\}]}.$$

As $N \to \infty$ we compute the stability equations for $A[Q]$ and obtain Q. It is known that the so called replica-symmetry (RS) ansatz $Q_{ab} = (1-Q)\delta_{ab} + Q$ leads to the correct result for the spherical perceptron, while it is not correct in the discrete case. Hence, the limit $Q \to 1^-$ identifies the point in which all the solutions collapse onto a single one (up to subextensive contributions) and it yields the critical point α_c [20].

In our model, this procedure (outlined in Appendix A) yields

$$\overline{F_\alpha(D_{\text{out}}|d_{\text{in}})} = \lim_{m\to 0} \frac{1}{Nm} \ln \int dQ\ d\hat{Q}\ e^{Nm[G_0(\hat{Q}) + G_1(Q, d_{\text{in}}, D_{\text{out}}) - iQ\hat{Q}/2]}. \tag{26}$$

The function G_0 depends on the model. For the spherical perceptron

$$G_0^{\text{spherical}}(\hat{Q}) = -\sqrt{i\hat{Q}} - i\hat{Q}/2, \tag{27}$$

while for the discrete model

$$G_0^{\text{discrete}}(\hat{Q}) = \ln 2 - i\hat{Q}/2 + \frac{1}{\sqrt{2\pi}} \int dx\ \exp\left(-\frac{1}{2}x^2\right) \ln\cosh\left[x\sqrt{i\hat{Q}}\right]. \tag{28}$$

The second function is common to the two models:

$$\begin{aligned} &G_1(Q, d_{\text{in}}, D_{\text{out}}) = \\ &\quad (1 - D_{\text{out}}) \left[\int \mathcal{D}_{d_{\text{in}}}(y, \bar{y}) \ln \int_{x>y;\ \bar{x}>\bar{y}} \mathcal{D}_{d_{\text{in}}}\left(\sqrt{\frac{Q}{1-Q}}x, \sqrt{\frac{Q}{1-Q}}\bar{x}\right)\right] \\ &\quad + D_{\text{out}} \left[\int \mathcal{D}_{d_{\text{in}}}(y, \bar{y}) \ln \int_{x>y;\ \bar{x}<\bar{y}} \mathcal{D}_{d_{\text{in}}}\left(\sqrt{\frac{Q}{1-Q}}x, \sqrt{\frac{Q}{1-Q}}\bar{x}\right)\right] \\ &\quad + D_{\text{out}} \ln D_{\text{out}} + (1 - D_{\text{out}}) \ln(1 - D_{\text{out}}), \end{aligned} \tag{29}$$

where

$$\mathcal{D}_{d_{\rm in}}(x,\bar{x}) = dx\; d\bar{x}\frac{1}{2\pi}\exp\left(-\frac{1}{2\sqrt{1-q^2}}\begin{bmatrix}x & \bar{x}\end{bmatrix}\begin{bmatrix}1 & -q\\ -q & 1\end{bmatrix}\begin{bmatrix}x\\ \bar{x}\end{bmatrix}\right), \tag{30}$$

with $q = 1-2d_{\rm in}$. It is worth noticing that in the limits $d_{\rm in}\to 0$ and $D_{\rm out}\to 0$, G_0 and G_1 reduce to Gardner and Derrida's RS formulas [18]. Finally, the saddle point equations are

$$-i\hat{Q}/2 = \frac{d}{dQ}G_1 \qquad -iQ/2 = \frac{d}{d\hat{Q}}G_0 \tag{31}$$

and are studied in Appendix B.

In order to compute the generalization capacity $\alpha_c(D_{\rm out}|d_{\rm in})$, we expand both sides of the equations and match the leading divergent terms. The result is:

$$\begin{aligned}\alpha_c(D_{\rm out}|d_{\rm in}) = \alpha_c\Bigg[&(1-D_{\rm out})\left(1+\frac{4}{\pi}\arctan\sqrt{\frac{d_{\rm in}}{1-d_{\rm in}}}\right)\\ &+D_{\rm out}\left(1+\frac{4}{\pi}\arctan\sqrt{\frac{1-d_{\rm in}}{d_{\rm in}}}\right)\Bigg]^{-1}\end{aligned} \tag{32}$$

with $\alpha_c = \alpha_c(0|0)$ being the RS capacity. It is $4/\pi$ for the discrete perceptron (which is known to be incorrect) and (correctly) 2 for the spherical perceptron. Hence, the RSB computation would be needed for the discrete case.

The distance-based capacity $\alpha_c(D_{\rm out}|d_{\rm in})$ is manifestly symmetric with respect to

$$(d_{\rm in}, D_{\rm out}) \mapsto (1-d_{\rm in}, 1-D_{\rm out})$$

2.2.3 Analytical and numerical phase diagram

The capacity computed above outlines a critical surface in the phase diagram defined by the parameters $d_{\rm in}, D_{\rm out}, \alpha$. Such surface is the boundary of the region within which the perceptron is expressive enough to realize the given task. A fixed value of $D_{\rm out}$ identifies a slice of the phase diagram (see Fig. 2). Let us fix $D_{\rm out} = 1$, meaning that the machine is required to assign the same output to $\boldsymbol{\xi}_\mu$ and $\bar{\boldsymbol{\xi}}_\mu$ for all μ. As expected, the capacity is a monotonically decreasing function of the distance, and one recovers the classic capacity $\alpha_{\rm c} = 2$ in the limit $d_{\rm in}\to 0$, where the two sets ξ and $\bar{\xi}$ coincide. Interestingly, $\alpha_c(0|d_{\rm in})$, displays a vertical tangent at $(0,\alpha_c)$. This means that, when the size of two input sets is close to the typical threshold value $\alpha_c N$, then their distance Ω typically goes to ∞ even for small values of their distance $d_{\rm in}$. In this limit, it is typically impossible for a perceptron to identify (give similar outputs to) two training sets even if they are very similar.

Another interesting limit is $d_{\rm in}\to 1$, whereby $\boldsymbol{\xi}_\mu \to -\bar{\boldsymbol{\xi}}_\mu$. In this limit, the problem at $D_{\rm out} = 1$ becomes equivalent to the linear separation of one-dimensional linear subspaces [15]. This problem can be solved marginally, meaning that its being solvable depends on the definition of the Heaviside θ function

in 0. More in general, such type of ambiguities is reflected by the fact that the limit $Q \to 1$ does not commute with those for $d_{\rm in} \to 0^+$ and $d_{\rm in} \to 1^-$, where the capacity is discontinuous. Some notable points are

$$\lim_{d_{\rm in}\to 0^+} \alpha_c(1|d_{\rm in}) = \frac{\alpha_c}{3} \tag{33}$$

$$\lim_{d_{\rm in}\to 1^-} \alpha_c(0|d_{\rm in}) = \frac{\alpha_c}{3} \tag{34}$$

$$\alpha_c(D_{\rm out} \neq 0|0) = \alpha_c(D_{\rm out} \neq 1|1) = 0. \tag{35}$$

These points show the existence of a critical value $\alpha = \alpha_c/3$ below which any degree of generalization is possible. More specifically, from (33) and (35), we see that, below this value, it is typically possible to distinguish arbitrarily similar input sets, unless they are identical. On the other hand, from (34) and (35), we see that it is possible to identify input sets which are arbitrarily different (unless they are exactly anti-parallel).

Overall, the behavior encoded by the phase diagram at fixed $D_{\rm out}$ can be expressed equivalently by fixing a value of the load α, and asking what is the critical $d_{\rm in}$ such that $\alpha_c(D_{\rm out}|d_{\rm in}) = \alpha$. At $D_{\rm out} = 0$, this allows one to interpret the distance-based capacity as a proxy for the maximum distance between two input sets that can be classified similarly. Equivalently, at $D_{\rm out} = 1$ this reasoning shows that the capacity encodes the minimum distance between two input sets that must be separated (i.e., classified differently).

Fig. 2 shows the agreement between the analytical formula (32) and numerical calculations, in the case of the spherical perceptron

with $N = 200$ input nodes. The network was trained with the usual perceptron algorithm, and the capacity was measured in the following way. For fixed number of patterns p, we tested whether the algorithm converged to a solution, for 1000 independent instances of the inputs ξ and $\bar{\xi}$ (at fixed input and output distances). The capacity was then computed as p^*/N, where p^* is the smallest value of p corresponding to which the algorithm failed to converge for at least half of the instances [31].

3 Discussion and Conclusions

The distance-based approach to generalization proposed here is complementary to existing ones, and quantifies how a network can recognize input sets having a prescribed degree of similarity to the training set. This approach corresponds to treating generalization as a task-dependent feature, rather than an absolute property of a neural network. In this framework, a network generalizes well if, given a training set, a typical network configuration assigns similar outputs to input patterns with fixed dissimilarity.

In order to demonstrate the applicability of our approach, we have performed explicit calculations for the perceptron, considering uncorrelated random inputs,

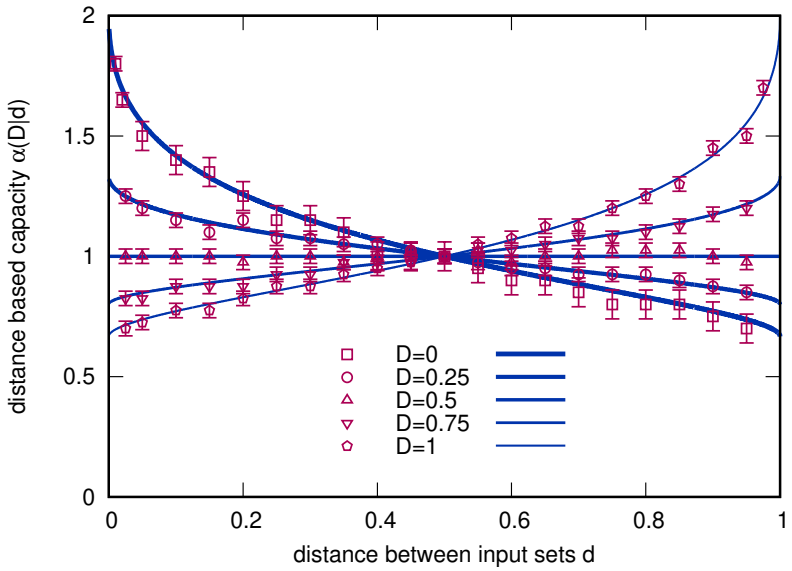

Figure 2: Distance-based capacity of a perceptron. The analytical prediction (solid lines) for the distance-based capacity, Eq. (32), is compared to numerical calculations (symbols), for the spherical perceptron. The five symbols correspond to the different values of $D_{\rm out}$ reported in the legend. The lowest capacity, $\alpha = 2/3$, is achieved by $\alpha_{\rm c}(D_{\rm out} = 0|d_{\rm in} = 1)$ and $\alpha_{\rm c}(D_{\rm out} = 1|d_{\rm in} = 0)$.

with the standard Hamming distance. Our calculations lead to the distance-based generalization capacity $\alpha_c(D_{\rm out}|d_{\rm in})$. The resulting phase diagram indicates that generalization is possible at any distance, but with decreasing critical capacity. The critical capacity has steeper drops close to the minimal and maximal distances, while showing slower decrease for intermediate increasing distances between the sets of patterns. The formula we have obtained for the critical line, Eq. (32), is formally equivalent to the one computed in a classic study [32] with completely different motivations. The special case $D_{\rm out} = 0$ was very recently rediscovered in a completely different context [33, 15], where it represents the capacity of a perceptron trained to discriminate segments. We surmise that the statistical-mechanics literature on neural networks, spanning the last 40 years, is replete with technical results awaiting to be rediscovered and reinterpreted in more contemporary settings.

Motivated by the analytical and numerical results reported here, we propose that averages computed by imposing a relation between members of the input set [such as the Gardner volume in Eq. (17)] are proxies for the dataset-dependent generalization capabilities of a machine. This is inspired by statistical learning theory, where the (dataset-*independent*) generalization error can be related to measures of the expressiveness of a given model, such as the Vapnik-Chervonenkis dimension, or other complexity measures. The results presented

above are a first step towards identifying a useful measure of dataset-dependent expressivity. However, we remark a limitation of the computations we have performed here. Averages such as that defined in Eq. (20) are computed by treating ξ and $\bar{\xi}$ on the same ground, meaning that we ask that the machine learn both input sets at the same time. In order to make quantitative contact with the conventional definition of generalization, the averaging procedure should be modified to have the machine learn ξ alone, and then test its results on $\bar{\xi}$. This is a challenging and stimulating problem for future work.

The concept of distance-based capacity is general and may be useful in numerical experiments with several kinds of neural networks. Rigorous and computationally feasible definitions of distances for real-world objects are important and challenging, as is well-known, for example, for images [34]. We have chosen here a simple instance of pattern distance but we expect that applying our approach using different system-tailored choices of distance metrics could reveal many aspects of how even complex neural networks operate. The distance-based approach may also be useful for gaining a better understanding of the phenomenon of catastrophic forgetting [35], i.e., the quick increase in the number of errors done on a first training set after the network is trained on a second set.

Finally, we comment briefly about a potential application of $F(D_{\rm out}|d_{\rm in})$ to probe the landscape of synaptic solutions. The existence of difference classes of solutions has been established in the case of the discrete simple perceptron. While dominant solutions are isolated, there exist clustered subdominant solutions, within a certain range [21, 22, 23]. Suppose the numbers $N_{\rm d}$ and $N_{\rm s}$ of dominant and subdominant solutions are exponentially large in the size N, with different rates $\Sigma_{\rm d}$ and $\Sigma_{\rm s}$ such that $\Sigma_{\rm d} > \Sigma_{\rm s}$. Then the subdominant class has zero measure in the thermodynamic limit, and is therefore "invisible" to conventional statistical averaging. Now consider the $F(0|d)$ solutions to the problem of learning two input sets ξ and $\bar{\xi}$ at distance $d_{\rm in}$. $\Sigma_{\rm d}$ and $\Sigma_{\rm s}$, as functions of $d_{\rm in}$, could cross each other at some value $d_{\rm in} = \tilde{d}_{\rm in}$, so that the subdominant set becomes dominant beyond $\tilde{d}_{\rm in}$. This would be signaled by a discontinuity in $F(0|d_{\rm in})$:

$$F(0|d_{\rm in}) = \lim_{N\to\infty} \frac{1}{N} \ln \left[\exp N\Sigma_{\rm d}(d_{\rm in}) + \exp N\Sigma_{\rm s}(d_{\rm in})\right] = \max[\Sigma_{\rm d}(d_{\rm in}), \Sigma_{\rm s}(d_{\rm in})]. \tag{36}$$

We speculate that this may be the case if dominant solutions are isolated and more sensitive to small input differences than clustered ones. More in general, non-analytic points of $F(D_{\rm out}|d_{\rm in})$ may reveal a non trivial landscape of solutions, highlighting the presence of different synaptic classes.

Acknowledgements

We thank Carlo Lucibello for feedback on the manuscript, and Sergio Caracciolo, Enrico Malatesta and Andrea di Gioacchino for useful discussions.

A Replica computation

In this section we will compute $F(D_{\text{out}}|d_{\text{in}})$ with the replica formalism:

$$F(D_{\text{out}}|d_{\text{in}}) = \lim_{m\to 0} \frac{1}{Nm} \ln \int P_{d_{\text{in}}}(\xi,\bar{\xi}) \prod_{a=1}^{m} \sum_{\{\epsilon_a^\mu=\pm 1\}} \prod_{a=1}^{m} \delta\left(\sum_{\mu=1}^{p} \epsilon_\mu^a - p(1-2D_{\text{out}})\right)$$
$$\int \prod_{a=1}^{m} dP(W^a) \prod_{\mu=1}^{p} \prod_{a=1}^{m} \Theta\left(W^a\cdot\xi_\mu\right)\ \Theta\left(\epsilon_\mu^a\ W^a\cdot\bar{\xi}_\mu\right) \tag{37}$$

We can introduce the parameters Q_{ab} and $\hat{Q}_{ab}$ by inserting the identy

$$1 = \int \prod_{a<b} dQ_{ab}\delta(Q_{ab} - W_a\cdot W_b/N) = N^m \int \prod_{a<b} dQ_{ab}\ d\hat{Q}_{ab}\ e^{iN\hat{Q}_{ab}\ (Q_{ab}-W_a\cdot W_b/N)}$$

in the integral $\int \prod_{a=1}^{m} dP(W_a)$. With this choice, $F(D_{\text{out}}|d_{\text{in}})$ can be rewritten as

$$F(D_{\text{out}}|d_{\text{in}}) = \lim_{m\to 0} \frac{1}{Nm} \ln \int \prod_{a<b} dQ_{ab}\ d\hat{Q}_{ab}\ e^{mN\left[i\sum_{a<b} Q_{ab}\hat{Q}_{ab} + G_0(\{\hat{Q}_{ab}\}) + G_1(\{Q_{ab}\},D_{\text{out}},d_{\text{in}})\right]} \tag{38}$$

with

$$G_0(\{\hat{Q}_{ab}\}) = \frac{1}{mN} \ln \int \prod_{a=1}^{m} dP(W_a)\ e^{-i\sum_{a<b}\hat{Q}_{ab}\ W_a\cdot W_b} \tag{39}$$

and

$$G_1(\{Q_{ab}\},D_{\text{out}},d_{\text{in}}) = \frac{1}{mN} \ln \int P_{d_{\text{in}}}(\xi,\bar{\xi}) \sum_{\{\epsilon_a^\mu=\pm 1\}} \prod_{a=1}^{m} \delta\left(\sum_{\mu=1}^{p} \epsilon_\mu^a - p(1-2D_{\text{out}})\right) \times$$
$$\times \prod_{\mu=1}^{p} \prod_{a=1}^{m} \Theta\left(W^a\cdot\xi_\mu\right)\ \Theta\left(\epsilon_\mu^a\ W^a\cdot\bar{\xi}_\mu\right) \tag{40}$$

The function G_1 has two properties:

- it depends on $\{W_a\}$ only though the overlap $W_a\cdot W_b/N = Q_{ab}$
- it can be rewritten in terms of Gaussian varibles $x_a^\mu = W_a\cdot\xi_\mu/\sqrt{N}$.

Both these properties can be deduced from the joint distribution of the auxiliary variables $\{x_a^\mu,\bar{x}_a^\mu\}_{a=1,\dots,m}$:

$$P_q^\mu(\{x_a^\mu,\bar{x}_a^\mu\}) := \frac{1}{P(d_{\text{in}})} P_{d_{\text{in}}}(\xi_\mu,\bar{\xi}_\mu) \prod_{a=1}^{m} \delta(x_a^\mu - W_a\cdot\xi^\mu/\sqrt{N})\ \delta(\bar{x}_a^\mu - W_a\cdot\bar{\xi}^\mu/\sqrt{N})$$
$$\underset{N\to\infty}{\to} \frac{1}{(2\pi)^m\sqrt{\det M(Q,q)}} \exp\left(-\frac{1}{2}\begin{bmatrix}x^\mu\\ \bar{x}^\mu\end{bmatrix}^t M^{-1}(Q,q)\begin{bmatrix}x^\mu\\ \bar{x}^\mu\end{bmatrix}\right) \tag{41}$$

with

$$x = \begin{bmatrix} x_1^\mu \\ ... \\ x_m^\mu \end{bmatrix} \qquad \bar{x} = \begin{bmatrix} \bar{x}_1^\mu \\ ... \\ \bar{x}_m^\mu \end{bmatrix} \qquad M(Q,q) = \begin{bmatrix} Q & qQ \\ qQ & Q \end{bmatrix} \qquad q = 1 - 2d_{\text{in}} \tag{42}$$

and Q being a shorthand notation for the matrix Q_{ab}. (41) only holds in the thermodynamic limit $N \to \infty$: a key passage consists in expanding $\ln\cosh(y/\sqrt{N}) \sim -\frac{y^2}{2N}$ for some finte y. Therefore

$$G_1(\{Q_{ab}\}, D_{\text{out}}, d_{\text{in}}) = \frac{1}{N} \ln \sum_{\{\epsilon_a^\mu = \pm 1\}} \prod_{a=1}^{m} \delta\left(\sum_{\mu=1}^{p} \epsilon_\mu^a - p(1-2D_{\text{out}})\right) \prod_{\mu=1}^{p} A_\mu(\{\epsilon_a^\mu\}, Q, q) \tag{43}$$

with

$$A_\mu(\{\epsilon_a^\mu\}, Q, q) = \int \prod_{a=1}^{m} dx_a^\mu \; d\bar{x}_a^\mu \; P_q^\mu(\{x_a^\mu, \bar{x}_a^\mu\}, Q) \; \prod_{a=1}^{m} \Theta(x_\mu^a) \; \Theta(\epsilon_a^\mu \; \bar{x}_\mu^a). \tag{44}$$

We can now introduce the RS ansatz

$$Q_{ab} = Q + (1-Q)\delta_{ab}. \tag{45}$$

Under this assumption, we can compute G_0 and G_1. The computation of G_0 is straightforward. In the discrete case,

$$G_0 = \ln\left\{ 2^m e^{-im\hat{Q}/2} \frac{1}{\sqrt{2\pi i\hat{Q}}} \int dx \; \exp\left(-\frac{1}{2i\hat{Q}}x^2 + m\ln\cosh(x)\right)\right\}. \tag{46}$$

If we expand it for small m, we get

$$G_0 = m\left\{\ln 2 - i\hat{Q}/2 + \frac{1}{\sqrt{2\pi}} \int dx \; \exp\left(-\frac{1}{2}x^2\right) \ln\cosh\left(x\sqrt{i\hat{Q}}\right)\right\} + o(m). \tag{47}$$

In the spherical case,

$$G_0 = -imJ - (1/2)\ln\left(\frac{iJ + (m-1)i\hat{Q}/2}{(iJ - i\hat{Q}/2)^{1-m}}\right) \tag{48}$$

with

$$iJ = \frac{1}{2}(\sqrt{i\hat{Q}} + i\hat{Q}).$$

For small m

$$G_0(\hat{Q}) = -m\left\{\sqrt{i\hat{Q}} + i\hat{Q}/2\right\}. \tag{49}$$

We can compute G_1 from (43). After some formal manipulations, the functions A_μ, as given by (44), can be rewritten as

$$A_\mu(D_\mu,Q,q)=\int \mathcal{D}_d(y,\bar y)\left[\int_{x>y;\ \bar x<\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right]^{mD_\mu}\times$$
$$\times\left[\int_{x>y;\ \bar x>\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right]^{m(1-D_\mu)} \tag{50}$$

with

$$\mathcal{D}_d(x,\bar x)=dx\ d\bar x\frac{1}{2\pi}\exp\left(-\frac{1}{2\sqrt{1-q^2}}\begin{bmatrix}x & \bar x\end{bmatrix}\begin{bmatrix}1 & -q\\ -q & 1\end{bmatrix}\begin{bmatrix}x\\ \bar x\end{bmatrix}\right) \tag{51}$$

and

$$m(1-2D_\mu)=\sum_{a=1}^{m}\epsilon_\mu^a. \tag{52}$$

For small m

$$A_\mu(D_\mu,Q,q)=\int \mathcal{D}_d(y,\bar y)\left[mD_\mu\ln\int_{x>y;\ \bar x<\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right.$$
$$\left.+\,m(1-D_\mu)\ln\int_{x>y;\ \bar x>\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right] \tag{53}$$

up to $o(m)$. Now, let us observe that

$$\sum_{\mu=1}^{p}mD_\mu=\frac{mp}{2}-\frac{1}{2}\sum_{\mu=1}^{p}\sum_{a=1}^{m}\epsilon_\mu^a=D_{\rm out}\,p\,m. \tag{54}$$

If we combine (54), (53) and (43), we get (29)

$$m\alpha G_1(Q,d,D_{\rm out})=m\alpha(1-D_{\rm out})\left[\int\mathcal{D}_d(y,\bar y)\ln\int_{x>y;\ \bar x>\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right]$$
$$+\,m\alpha D_{\rm out}\left[\int\mathcal{D}_d(y,\bar y)\ln\int_{x>y;\ \bar x<\bar y}\mathcal{D}_d\left(\sqrt{\frac{Q}{1-Q}}x,\sqrt{\frac{Q}{1-Q}}\bar x\right)\right]$$
$$+\,m\alpha[D_{\rm out}\ln D_{\rm out}+(1-D_{\rm out})\ln(1-D_{\rm out})]+o(m). \tag{55}$$

For the sake of simplicity, we can rewrite (55) as

$$G_1(Q,d,D_{\rm out})=:(1-D_{\rm out})A_+(Q,d,D_{\rm out})+D_{\rm out}A_-(Q,d,D_{\rm out})+$$
$$+\,D_{\rm out}\ln D_{\rm out}+(1-D_{\rm out})\ln(1-D_{\rm out}). \tag{56}$$

B Saddle point equations and derivation of the critical capacity

From (38) we read the saddle point equations:

$$iQ_{ab} = \frac{d}{d\hat{Q}_{ab}} G_0 \tag{57}$$

$$i\hat{Q}_{ab} = \frac{d}{dQ_{ab}} G_1. \tag{58}$$

If we use the RS ansatz (45), then (57) and (58) become

$$-imQ/2 = \frac{d}{d\hat{Q}} G_0 \tag{59}$$

$$-im\hat{Q}/2 = \frac{d}{dQ} G_1 \tag{60}$$

as $m \to 0$. Eq. (60) can be rewritten as (see (56) for the notation)

$$-\frac{i}{2}\hat{Q} = \alpha \frac{d}{dQ}[DA_- + (1-D)A_+]. \tag{61}$$

Eq. (59) for the discrete model is

$$-\frac{1}{2}Q = -1/2 + \frac{1}{\sqrt{8\pi i\hat{Q}}} \int dx\, x \exp\left(-\frac{1}{2}x^2\right) \tanh\left(x\sqrt{i\hat{Q}}\right) \tag{62}$$

while for the spherical model

$$-\frac{1}{2}Q = -\frac{1}{2}\left(1 + 1/\sqrt{i\hat{Q}}\right). \tag{63}$$

From the previous equations, we can write

$$\alpha(d_{\text{in}}, D_{\text{out}}, Q) = \frac{-i\hat{Q}(Q)}{2\frac{d}{dQ}[D_{\text{out}}A_-(d_{\text{in}}, Q) + (1-D_{\text{out}})A_+(d_{\text{in}}, Q)]}. \tag{64}$$

In order to get the critcal capacity, we should evaluate $\alpha(d_{\text{in}}, D_{\text{out}}, Q)$ at $Q = 1$:

$$\alpha_c(D_{\text{out}}|d_{\text{in}}) = \lim_{Q\to 1^-} \frac{-i\hat{Q}(Q)}{2\frac{d}{dQ}[D_{\text{out}}A_-(d_{\text{in}}, Q) + (1-D_{\text{out}})A_+(d_{\text{in}}, Q)]}. \tag{65}$$

However, the quantities appearing in the previous equation are neither defined nor analytical at $Q = 1$. For this reason we have to extract the leading divergencies around $Q = 1^-$. We have to expand $D_{\text{out}}A_-(d_{\text{in}}, Q) + (1-D_{\text{out}})A_+(d_{\text{in}}, Q)$ first. For this purpose, it is convenient to perform a change of variables $S =$

$(y+\bar{y})/\sqrt{2}$, $D=(y-\bar{y})/\sqrt{2}$ (not to be confused with the output difference), $s=(x+\bar{x})/\sqrt{2}$ and write:

$$A_- = \frac{1}{2\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}S^2-\frac{1}{2}D^2\right)\ln\frac{1}{\sqrt{2\pi}}\times$$
$$\times\left\{\int_0^\infty ds\, \exp\left(-\frac{1}{2}\frac{Q}{1-Q}(s-S)^2\right)\operatorname{erfc}\left(\left[\sqrt{\frac{Q}{1-Q}}\left[D+\sqrt{\frac{1+q}{1-q}}s\right]\right]\right)\right.$$
$$\left.+\int_0^\infty ds\, \exp\left(-\frac{1}{2}\frac{Q}{1-Q}(s+S)^2\right)\operatorname{erfc}\left(\left[\sqrt{\frac{Q}{1-Q}}\left[D+\sqrt{\frac{1+q}{1-q}}s\right]\right]\right)\right\} \tag{66}$$

$$A^+ = \frac{1}{2\pi}\int dS\, dD\, \exp\left(-\frac{1}{2}S^2-\frac{1}{2}D^2\right)\ln\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-Q}{Q}}\int_0^\infty ds\, \exp\left(-\frac{1}{2}\frac{Q}{1-Q}(s+S)^2\right)$$
$$\left\{\operatorname{erfc}\left(\sqrt{\frac{Q}{1-Q}}\left[D-\sqrt{\frac{1+q}{1-q}}s\right]\right)-\operatorname{erfc}\left(\sqrt{\frac{Q}{1-Q}}\left[D+\sqrt{\frac{1+q}{1-q}}s\right]\right)\right\} \tag{67}$$

where we have used the definition $\operatorname{erfc}(y)=\frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-x^2/2}$. We can use the following limit in order to perform the expansion:

$$\lim_{\epsilon\to 0^+}\epsilon\ln\int_I dx\, e^{-ax^2/\epsilon}\operatorname{erfc}((bx-c)/\sqrt{\epsilon})=\min_{x\in I}(ax^2+\Theta(bx-c)(bx-c)^2). \tag{68}$$

If we set $\epsilon=1-Q$ in (66) and (67), then we obtain that

$$A_\pm(Q\to 1)\sim\frac{1}{1-Q}\int\mathcal{D}x\,\mathcal{D}y\,\sum_j P_j^\pm(x,y)\,\chi_{\Omega_j}, \tag{69}$$

where $\mathcal{D}x=dx\,\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, χ is the characteristic function, P_js are second degree polynomials and Ω_js are disjoint sets such that $\cup_j\Omega_j=\mathbb{R}^2$. The explicit computation is a bit lengthy but straightforward and the result is

$$A_\pm(Q\to 1)\sim\frac{1}{1-Q}\left\{1+\frac{4}{\pi}\arctan\left[\left(\frac{1-q}{1+q}\right)^{\pm 1/2}\right]\right\}. \tag{70}$$

However, while $A_\pm(Q\to 1)$ (and their derivatives) are finite at both $d_{\text{in}}=1$ and $d_{\text{in}}=0$, if we fix Q, then

$$\lim_{d_{\text{in}}\to 0^+}A_-=\lim_{d_{\text{in}}\to 1^-}A_+=-\infty \tag{71}$$

since

$$\lim_{\epsilon \to 0^+} \int dx \; e^{-ax^2} \; \text{erfc}(x/\epsilon - b) = 0.$$

Therefore

$$\lim_{d_{\text{in}} \to 1^-} \frac{-i\hat{Q}(Q)}{2\frac{d}{dQ}[D_{\text{out}} A_-(d_{\text{in}}, Q) + (1 - D_{\text{out}}) A_+(d_{\text{in}}, Q)]} = 0 \tag{72}$$

unless $D_{\text{out}} = 1$, and

$$\lim_{d_{\text{in}} \to 0^+} \frac{-i\hat{Q}(Q)}{2\frac{d}{dQ}[D_{\text{out}} A_-(d_{\text{in}}, Q) + (1 - D_{\text{out}}) A_+(d_{\text{in}}, Q)]} = 0 \tag{73}$$

unless $D_{\text{out}} = 0$. The conclusion is that the limits in $(D_{\text{out}}, d_{\text{in}})$ and Q do not commute. The consequence is the existence of a delta discontinuity. Finally, we can combine (72), (73), (70) and (65) into the distance-based capacity in (32).

References

[1] M. A. Gluck, C. E. Myers, and E. Mercado. *Learning and Memory: The Brain in Action*. Crane Library at the University of British Columbia, 2011.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[3] Charles H Martin and Michael W Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

[4] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45:6056–6091, Apr 1992.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[6] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, Sept 1999.

[7] P. Mehta, M. Bukov, C. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *arXiv preprint arXiv:1803.08823*, 2018.

[8] G. Friedland, A. Metere, and M. Krell. A practical approach to sizing neural networks. *arXiv preprint arXiv:1810.02328*, 2018.

[9] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

[10] Manfred Opper. Learning to generalize. *Frontiers of Life*, 3(part 2):763–775, 2001.

[11] Rémi Monasson and Riccardo Zecchina. Weight space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks. *Phys. Rev. Lett.*, 75:2432–2435, Sep 1995.

[12] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, 78(10):1568–1574, Oct 1990.

[13] Pietro Rotondo, Marco Cosentino Lagomarsino, and Marco Gherardi. Counting the learnable functions of structured data. *arXiv:1903.12021 [cond-mat.dis-nn]*, 2019.

[14] Andrea Mazzolini, Marco Gherardi, Michele Caselle, Marco Cosentino Lagomarsino, and Matteo Osella. Statistics of shared components in complex component systems. *Phys. Rev. X*, 8:021023, Apr 2018.

[15] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, 8:031003, Jul 2018.

[16] Andrea Mazzolini, Jacopo Grilli, Eleonora De Lazzari, Matteo Osella, Marco Cosentino Lagomarsino, and Marco Gherardi. Zipf and heaps laws from dependency structures in component systems. *Phys. Rev. E*, 98:012315, Jul 2018.

[17] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Phys. Rev. E*, 95:022117, Feb 2017.

[18] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, 1988.

[19] W K Theumann and R Erichsen Jr. Gardner-derrida neural networks with correlated patterns. *Journal of Physics A: Mathematical and General*, 24(10):L565, 1991.

[20] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[21] Haiping Huang and Yoshiyuki Kabashima. Origin of the computational hardness for learning with binary synapses. *Phys. Rev. E*, 90:052813, Nov 2014.

[22] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Phys. Rev. Lett.*, 115:128101, Sep 2015.

[23] Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016.

[24] R Erichsen and W K Thuemann. Optimal storage of a neural network model: a replica symmetry-breaking solution. *Journal of Physics A: Mathematical and General*, 26(2):L61, 1993.

[25] Rémi Monasson. Storage of spatially correlated patterns in autoassociative memories. *Journal de Physique I*, 3(5):1141–1152, 1993.

[26] R Monasson. Properties of neural networks storing spatially correlated patterns. *Journal of Physics A: Mathematical and General*, 25(13):3701, 1992.

[27] Y Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series*, 95(1):012001, 2008.

[28] Takashi Shinzato and Yoshiyuki Kabashima. Learning from correlated patterns by simple perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 42(1):015005, 2009.

[29] Osame Kinouchi and Nestor Caticha. Learning algorithm that gives the bayes generalization limit for perceptrons. *Phys. Rev. E*, 54:R54–R57, Jul 1996.

[30] Manfred Opper and David Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, 66:2677–2680, May 1991.

[31] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, June 1965.

[32] B Lopez, M Schroder, and M Opper. Storage of correlated patterns in a perceptron. *Journal of Physics A: Mathematical and General*, 28(16):L447, 1995.

[33] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Linear readout of object manifolds. *Phys. Rev. E*, 93:060301, Jun 2016.

[34] Vito Di Gesu and Valery Starovoitov. Distance-based functions for image comparison. *Pattern Recognition Letters*, 20(2):207 – 214, 1999.

[35] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109 – 165. Academic Press, 1989.