Demystification of Few-shot and One-shot Learning

1st Ivan Y. Tyukin
School of Mathematics and Actuarial Science
University of Leicester
and *Norwegian University of Science and Technology*
and *Saint-Petersburg State Electrotechnical University*

Leicester, United Kingdom
and Trondheim, Norway
and Saint-Petersburg, Russia
i.tyukin@le.ac.uk

2nd Alexander N. Gorban
School of Mathematics and Actuarial Science
University of Leicester
and *Lobachevsky University*

Leicester, United Kingdom
and Nizhny Novgorod, Russia
a.n.gorban@le.ac.uk

3rd Muhammad H. Alkhudaydi
School of Mathematics and Actuarial Science
University of Leicester

Leicester, United Kingdom
mhaa4@le.ac.uk

4th Qinghua Zhou
School of Informatics
University of Leicester

Leicester, United Kingdom
qz105@le.ac.uk

***Abstract*—Few-shot and one-shot learning have been the subject of active and intensive research in recent years, with mounting evidence pointing to successful implementation and exploitation of few-shot learning algorithms in practice. Classical statistical learning theories do not fully explain why few- or one-shot learning is at all possible since traditional generalisation bounds normally require large training and testing samples to be meaningful. This sharply contrasts with numerous examples of successful one- and few-shot learning systems and applications.**

In this work we present mathematical foundations for a theory of one-shot and few-shot learning and reveal conditions specifying when such learning schemes are likely to succeed. Our theory is based on intrinsic properties of high-dimensional spaces. We show that if the ambient or latent decision space of a learning machine is sufficiently high-dimensional than a large class of objects in this space can indeed be easily learned from few examples provided that certain data non-concentration conditions are met.

***Index Terms*—Few-shot learning, one-shot learning, generalisation, stochastic separation theorems**

NOTATION

- $\mathbb{R}$ denotes the field of real numbers, $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} |\ x \geq 0\}$, and $\mathbb{R}^n$ stands for the n-dimensional linear real vector space;
- $\mathbb{N}$ denotes the set of natural numbers;
- bold symbols $\boldsymbol{x} = (x_1, \ldots, x_n)$ will denote elements of $\mathbb{R}^n$;
- $(\boldsymbol{x}, \boldsymbol{y}) = \sum_k x_k y_k$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ is the standard Euclidean norm in $\mathbb{R}^n$;
- $\mathbb{B}_n$ denotes the unit ball in $\mathbb{R}^n$ centered at the origin:

$$\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n |\ \|\boldsymbol{x}\| \leq 1\};$$

- $\mathbb{B}_n(r, \boldsymbol{y})$ stands for the ball in $\mathbb{R}^n$ of radius $r > 0$ centered at $\boldsymbol{y}$:

$$\mathbb{B}_n(r, \boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}^n |\ \|\boldsymbol{x} - \boldsymbol{y}\| \leq r\};$$

The work was supported by the UKRI Alan Turing AI Acceleration Fellowship grant EP/V025295/1 and by the grant of the Ministry of Science and Higher Education of Russian Federation (Project No. 14.Y26.31.0022).

- V_n is the n-dimensional Lebesgue measure, and $V_n(\mathbb{B}_n)$ is the volume of unit n-ball;

I. INTRODUCTION

The fundamental question of learning from few examples is one of the fascinating and central questions in both the theory and practice of modern large-scale data-driven AI systems. These systems have many millions of adjustable parameters [1], whose numbers often exceed those of the datasets used in their training. And yet, performance of these large-scaled models trained on modestly-sized datasets in practical tasks is remarkably good [2].

Classical generalisation bounds stemming from the Vapnik-Chervonenkis theory [3] alone do not explain these successes due to their combinatorial and extremely conservative nature. What is even more striking, as has been demonstrated in [4], absolutely identical deep neural networks are capable of exhibiting both sides of the learning spectrum: to successfully generalise from meaningful training data and, at the same time, "memorise" random assignments of labels without any generalisation. Results like these motivate persistent ongoing inquiries into unreasonable effectiveness of modern deep learning models [5].

The phenomenon of few-shot learning is perhaps one of the most acute manifestations of this challenge. Various few-shot learning schemes and empirically successful algorithms and models such as matching [6] and prototypical networks [7] provide ample evidence that good generalisation may indeed occur in extreme settings with just few elements in the training set. The theory, however, which may explain why is this at all possible is lacking.

In this paper, we lay out mathematical foundations of such theory. We provide, for the first time, *formal statements* of different versions of the problem of few-shot learning and *present solutions* of these problems. These solutions are remarkably consistent with heuristic algorithms described in the current

literature [6], [7]. At the core of our approach are stochastic separation theorems [8], [9] linking high-dimensional geometry with the concentration of measure. In this work, we make an additional departure from the classical "fully agnostic" machine learning problem statement. In particular, we propose that a mild hypothesis on "compactness" of an object's/class's representation in the network's latent space, expressed as existence of a finite sub-cover of the object to be learned by n-balls not containing the origin, could hold the key to understanding and resolving the challenge of generalisation, few-shot, and single-shot learning.

The rest of the paper is organised as follows. In Section II we describe a general setting of the problem of few-shot learning considered in the paper and present its formal mathematical statements. Section III presents main mathematical results and their discussion, and Section IV provides a brief summary and conclusion.

II. Problem formulation

A. General setting

To set the scene for a more formal analysis, let us first outline key components of few-shot learning. In many relevant few-shot learning cases one would normally have an *existing system* with all its inputs, outputs, states and dependencies (potentially unknown) between these. This existing system would also operate in a specific regime (recognise a new person in a room, learn a new gesture, fix an error) which can be termed as an *operational situation*. Performance of the system in the task of learning in this situation is then assessed by some *evaluation procedure*.

Complexity of all processes presented in this rather generic picture could be extremely high. In modern large-scale AI and deep learning models, one of the major contributors to this complexity is an inherently and irreducibly *high dimensionality* of signals involved in the definition of the operational situation at hand. Defining meaningful probability spaces for such data is not a trivial task due to enormously large datasets required to gain appropriate knowledge and intuition. At the same time, as we will show later, this high dimensionality may hold the key to develop some understanding of the phenomenon of few-shot learning.

In order to reveal the link between dimensionality of the appropriate data and few-shot learning we will need to make some simplifying assumptions constraining the general setting above. These assumptions, however, would enable us to define the problem formally and focus on the most relevant elements of the general problem which are important for this contribution. In the next section we provide a formal, albeit simplified, description of the problem (II-B), formalise the problem of few-shot learning (II-C), and list some specific technical assumptions (II-D).

B. Background

Let $\mathcal{U}, \mathcal{U} \subset \mathbb{R}^d$ be the set of inputs modeling or representing objects of interests such as images, pieces of sound, or records in a database, and let $\mathcal{L}$ be the set of labels. Following classical statistical learning settings [10], [3], we suppose that for each element $\boldsymbol{u} \in \mathcal{U}$ there is an associated label $\ell \in \mathcal{L}$, and that the pairs $(\boldsymbol{u}_i, \ell_i) \in \mathcal{U} \times \mathcal{L}$, $i = 1, \ldots, N$, $N \in \mathbb{N}$ are observations drawn from some joint probability distribution. For convenience, we shall assume that there exist some corresponding distributions P_u and $P_{\ell|u}$ such that the joint distribution of $\boldsymbol{u}$ and ℓ is expressed as: $P_{\ell|u}(\ell|\boldsymbol{u})P_u(\boldsymbol{u})$.

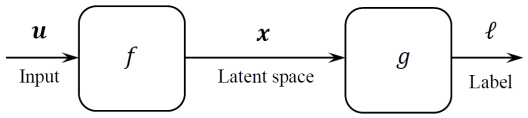

Fig. 1. Assumed input-output workflow of the classifier subject to few-shot learning tasks

To formally specify the problem of few-shot learning and its relevant variants, we need to determine a system that would be subjected to such learning. For the sake of simplicity, here we will assume that this system is a classifier. In general, however, this latter assumption may be dropped, and the problem of few-shot learning could be extended to much broader classes of AI systems.

Let $F : \mathcal{U} \to \mathcal{L}$ be such classifier assigning a unique label from the set $\mathcal{L}$ (the set defining all possible lables) to an element from $\mathcal{U}$. In what follows we shall assume that $F = g \circ f$ where

$$f : \mathcal{U} \to \mathcal{X}, \ \mathcal{X} \subset \mathbb{R}^n \tag{1}$$

defines the classifier's F latent space $\mathcal{X}$, and

$$g : \mathcal{X} \to \mathcal{L}$$

determines how the classifier F assigns a label to an input $\boldsymbol{u}$ having the the corresponding latent representation $\boldsymbol{x} = f(\boldsymbol{u})$. A diagram showing schematic representation of the classifier's workflow is shown in Fig. 1.

The above structure is very general and covers the majority of existing classification models. We are now ready for formal definitions of the relevant few-shot learning problems.

C. Few-shot learning problems

In what follows we consider two classes of few-shot learning problems: *learning new examples* from their single representation, and *learning a new class* from few examples. These problems have different uses and aims. The latter focuses primarily on *generalising* from a limited number of data points, whereas the former aims at *memorising* new data without destroying existing knowledge in the system.

1) Learning a finite number of new examples: We begin with the first version of the problem, where the task is to learn, or memorise, a given finite set. This task is formally introduced as Problem 1 below.

Problem 1 (Learning few examples): Consider a classifier F defined by (1), and let $\mathcal{U}_{\text{new}} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$, $k \in \mathbb{N}$, $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$, be a given finite set to be learned by F. Let $\ell_{\text{new}} \in \mathcal{L}$

be a label associated with the new set $\mathcal{U}_{\text{new}}$. Let p_e be a given positive number in the interval $(0, 1]$ determining the quality of learning.

Find an algorithm $\mathcal{A}(\mathcal{U}_{\text{new}})$ producing a function $g^* : \mathcal{X} \to \mathcal{L}$ such that

$$F(g^* \circ f(\boldsymbol{u})) = \ell_{\text{new}} \text{ for all } \boldsymbol{u} \in \mathcal{U}_{\text{new}} \tag{2}$$

and

$$P\big(F(g^* \circ f(\boldsymbol{u})) = F(g \circ f(\boldsymbol{u}))\big) \geq p_e \tag{3}$$

for $\boldsymbol{u}$ drawn from the distribution P_u.

2) Learning from an arbitrary finite number of examples: Let us now consider a different version of the problem where the system is to *learn a new class* from few examples. The key difference here from the case considered in Problem 1 is that we will no longer require that all new examples are memorised. Instead, we will request that *all* elements of the new class are assigned a correct label with some a-priori defined probability. At the same time, we will request that performance of the classifier on elements from other classes does not drop below a given predefined and acceptable level.

Extending our earlier conventions, we will suppose that the new class can be described by a corresponding probability distribution P_{new} and will be associated with a new label ℓ_{new}. Formal statement of this task is provided in Problem 2.

Problem 2 (Learning from few examples): Consider a classifier F defined by (1), and let $\mathcal{U}_{\text{new}} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$, $k \in \mathbb{N}$, $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$, be a finite independent and identically distributed (i.i.d.) sample from a distribution P_{new}, and $\ell_{\text{new}} \in \mathcal{L}$ be a corresponding new label to be associated with the elements drawn from P_{new}. Let p_e, p_n be given positive numbers in the interval $(0, 1]$ determining the quality of learning.

Find an algorithm $\mathcal{A}(\mathcal{U}_{\text{new}})$ producing a function $g^* : \mathcal{X} \to \mathcal{L}$ such that

$$P\big(F(g^* \circ f(\boldsymbol{u})) = \ell_{\text{new}}\big) \geq p_n \tag{4}$$

for $\boldsymbol{u}$ drawn from P_{new}, and

$$P\big(F(g^* \circ f(\boldsymbol{u})) = F(g \circ f(\boldsymbol{u}))\big) \geq p_e \tag{5}$$

for $\boldsymbol{u}$ drawn from the distribution P_u.

Remark 1: Note that Problems 1, 2 do not rely upon standard relationships between expected and empirical risks to characterise generalisation and learning. Instead, they impose stronger requirements: lower bounds on probabilities of success.

These stronger requirements have clear practical benefits in terms of understanding limitations and capabilities of few-shot learning algorithms $\mathcal{A}$. Potential downsides, however, are that knowledge of some general properties of the data distributions (support, non-degeneracy, etc) may be needed to guarantee that these stronger requirements could be met.

D. Assumptions

In agreement with existing literature on few-shot learning [6], [7], we will primarily be dealing with representations $\boldsymbol{x} = f(\boldsymbol{u})$ of inputs $\boldsymbol{u} \in \mathcal{U}$ in the system's latent space $\mathcal{X}$ as opposed to working directly with $\mathcal{U}$ (see Fig. 1 for a diagram of the workflow). We will hence assume that the distributions P_u, $P_{\ell|u}$, P_{new}, and the function f in (1) – (5) induce their corresponding distributions P_x, $P_{\ell|x}$, $P_{\text{new},x}$ in the system's latent space $\mathcal{X}$.

We will further assume that distributions P_x, $P_{\text{new},x}$ are supported on some balls in $\mathbb{R}^n$ and admit probability density functions satisfying some non-degeneracy constraints. Formally these requirements are formulated in Assumptions 1, 2.

Assumption 1: The probability density function p_x associated with P_x exists, is defined on the unit ball $\mathbb{B}_n$, and there exist constants $C_x, r > 0$ such that

$$p_x(\boldsymbol{x}) \leq \frac{C_x}{V_n(\mathbb{B}_n)} r^n.$$

Assumption 2: The probability density function $p_{\text{new},x}$ associated with $P_{\text{new},x}$ exists, is defined on a ball $\mathbb{B}_n(v, \boldsymbol{c})$, and there exist constants $C_{\text{new},x}, \rho > 0$ such that

$$p_{\text{new},x}(\boldsymbol{x}) \leq \frac{C_{\text{new},x}}{V_n(\mathbb{B}_n)} \rho^n.$$

In the next section we present main theoretical findings and quantifying success of few- and one-shot learning schemes. These results join together various ideas presented in earlier works [11], [12], [13], [14], [15], [16], and reveal intrinsic links between data dimensionality, partial knowledge about data models, and generalisation bounds.

III. Main Results

A. Learning an arbitrary finite number of examples

Our first result concerns Problem 1 and is formally expressed in Theorem 1 below

Theorem 1: [Learning few examples] Consider a classifier F defined by (1), and let $\mathcal{U}_{\text{new}} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$, $k \in \mathbb{N}$, $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$, be a finite set, and $\ell_{\text{new}} \in \mathcal{L}$ be a corresponding new label to be associated with the elements from this new set.

Let $\mathcal{Y} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, $\boldsymbol{x}_i = f(\boldsymbol{u}_i)$, $i = 1, \ldots, k$ be a representation of the set $\mathcal{U}_{\text{new}}$ in the classifier's latent space,

$$\bar{\boldsymbol{x}} = \frac{1}{k} \sum_i \boldsymbol{x}_i,$$

be the empirical mean of the representation with

$$(\bar{\boldsymbol{x}}, \boldsymbol{x}_i) \geq 0 \text{ for all } \boldsymbol{x}_i \in \mathcal{Y},$$

and let Assumption 1 hold.

Then the map

$$g^* : \; g^*(\boldsymbol{x}) = \begin{cases} \ell_{\text{new}}, & \text{if } \left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}\right) - \theta \geq 0 \\ g(\boldsymbol{x}), & \text{otherwise} \end{cases} \tag{6}$$

parameterised by

$$\theta = \min_{i \in \{1, \ldots, k\}} \left\{ \left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}_i \right) \right\}$$

is a solution of Problem 1 with

$$p_e = 1 - \frac{C_x}{2} \left[r \left(1 - \theta^2\right)^{1/2} \right]^n. \tag{7}$$

Proof of Theorem 1. According to the definition of the map g^* and the fact that $\boldsymbol{x}_i = f(\boldsymbol{u}_i)$,

$$\left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}_i\right) - \theta \geq 0$$

and as a result

$$F(g^* \cdot f(\boldsymbol{u}_i)) = \ell_{\text{new}}$$

for any $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$.

Let $\boldsymbol{u}$ be drawn from the distribution P_u. This vector has a latent representation $\boldsymbol{x} = f(\boldsymbol{u})$ and a corresponding induced distribution P_x satisfying Assumption 1. Let

$$\mathcal{C}_n(\boldsymbol{z}, \theta) = \left\{ \boldsymbol{x} \in \mathbb{B}_n \,\middle|\, \left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, \boldsymbol{x}\right) - \theta \geq 0 \right\}.$$

Then

$$P(F(g^* \circ f(\boldsymbol{u})) = \ell_{\text{new}}) = \int_{\mathcal{C}_n(\bar{\boldsymbol{x}}, \theta)} p_x(\boldsymbol{x}) d\boldsymbol{x}$$
$$\leq \frac{C_x}{V_n(\mathbb{B}_n)} r^n \int_{\mathcal{C}_n(\bar{\boldsymbol{x}}, \theta)} d\boldsymbol{x} \leq \frac{C_x}{2} r^n [(1-\theta^2)^{1/2}]^n.$$

The statement now follows. □

Remark 2: Note that if $r(1-\theta^2)^{1/2} < 1$ then the bound p_e approaches 1 exponentially fast as n grows. This implies that learning a single or few examples can be efficiently accomplished by an exceptionally simple map (6).

Performance of this learning scheme depends on the values of r and θ. The closer the value of θ is to 1, however, the broader the range of r for which solution (6) of Problem 1 is appropriate.

Remark 3: One can easily verify that few-shot learning scheme (6) assigns the label ℓ_{new} to all convex combinations of $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$. In this respect, the entire convex hull of $\mathcal{U}_{\text{new}}$ is learnt by (6).

In the next subsection we will show that, under appropriate assumptions, learning schemes which are very similar to (6) have a capacity to generalise beyond finite sets and their convex hulls from just few examples.

B. Learning from few examples

Let us now turn attention to Problem 2. Our main theoretical statement specifying a simple solution of this problem is presented in Theorem 2. Similarly to Theorem 1, we show that performance of the proposed scheme to learn from k examples is closely related to 1) dimension n of the classifier's latent space and 2) non-degeneracy of probability distributions of the inputs' representations in that space.

The theorem is largely based on Lemmas 1, 2 which we present below.

Lemma 1: Let $\mathcal{Y} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ be a set of k i.i.d. random vectors drawn from a distribution satisfying Assumption 2, and let $\delta, \varepsilon \in (0,1)$. Consider event A_1:

$$A_1: \ |(\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{x}_j - \boldsymbol{c})| \leq \delta v, \ \forall \ i \neq j \tag{8}$$

and event A_2:

$$A_2: \ \|\boldsymbol{x}_i - \boldsymbol{c}\| \geq (1-\varepsilon) v \ \forall \ i. \tag{9}$$

Then

$$P(A_1) \geq 1 - C_{\text{new}} \frac{k(k-1)}{2} \left[\rho v (1-\delta^2)^{1/2}\right]^n, \tag{10}$$

and

$$P\left(A_1 \wedge A_2\right) \geq$$
$$1 - C_{\text{new}} k [\rho v (1-\varepsilon)]^n - C_{\text{new}} \frac{k(k-1)}{2} \left[\rho v (1-\delta^2)^{1/2}\right]^n. \tag{11}$$

Proof of Lemma 1. Let us denote $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{c}$. Consider events

$$E_1: \ \left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_2\right)\right| > \delta$$
$$E_2: \ \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_3\right)\right| > \delta\right] \vee \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_2}{\|\tilde{\boldsymbol{x}}_2\|}, \tilde{\boldsymbol{x}}_3\right)\right| > \delta\right]$$
$$\vdots \qquad\qquad \vdots$$
$$E_{k-1}: \ \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_k\right)\right| > \delta\right] \vee \cdots \vee \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_{k-1}}{\|\tilde{\boldsymbol{x}}_{k-1}\|}, \tilde{\boldsymbol{x}}_k\right)\right| > \delta\right]$$

$$B_1: \|\tilde{\boldsymbol{x}}_1\| < (1-\varepsilon)v$$
$$\vdots \qquad\qquad \vdots$$
$$B_k: \|\tilde{\boldsymbol{x}}_k\| < (1-\varepsilon)v$$

Let

$$\mathcal{C}_n(\boldsymbol{z}, \boldsymbol{c}, v, \delta) = \left\{ \boldsymbol{x} \in \mathbb{B}_n(v, \boldsymbol{c}) \,\middle|\, \left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, \boldsymbol{x} - \boldsymbol{c}\right) > \delta \right\}.$$

According to Assumption 2 and the fact that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are drawn independently from the same distribution, the probability that event E_1 occurs can be bounded from above as

$$\begin{aligned} P(E_1) = &\int_{\mathcal{C}_n(\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)} p_{\text{new},x}(\boldsymbol{x}) d\boldsymbol{x} \\ &+ \int_{\mathcal{C}_n(-\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)} p_{\text{new},x}(\boldsymbol{x}) d\boldsymbol{x} \\ < &\frac{C_{\text{new},x}\rho^n}{V_n(\mathbb{B}_n)} \int_{\mathcal{C}_n(\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)} d\boldsymbol{x} \\ &+ \frac{C_{\text{new},x}\rho^n}{V_n(\mathbb{B}_n)} \int_{\mathcal{C}_n(-\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)} d\boldsymbol{x} \\ = &\frac{C_{\text{new},x}\rho^n}{V_n(\mathbb{B}_n)} 2 V_n(\mathcal{C}_n(\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)). \end{aligned} \tag{12}$$

Observe that

$$2V_n\big(\mathcal{C}_n(\tilde{\boldsymbol{x}}_1, \boldsymbol{c}, v, \delta)\big) \leq V_n(\mathbb{B}_n)[v(1-\delta^2)^{1/2}]^n. \tag{13}$$

Combining (12), (13) we obtain:

$$P(E_1) < C_{\text{new},x}[\rho v (1-\delta^2)^{1/2}]^n. \tag{14}$$

Recall that for any events $A_1, \ldots, A_k$ the following probability union bound holds true (also known as the Boole's inequality):

$$P(\bigcup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} P(A_i). \tag{15}$$

Hence

$$P(E_2) = P\left(\left[\left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_3\right)\right| > \delta\right] \vee \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_2}{\|\tilde{\boldsymbol{x}}_2\|}, \tilde{\boldsymbol{x}}_3\right)\right| > \delta\right]\right)$$
$$\leq \sum_{i=1}^{2} P\left(\left|\left(\frac{\tilde{\boldsymbol{x}}_i}{\|\tilde{\boldsymbol{x}}_i\|}, \tilde{\boldsymbol{x}}_3\right)\right| > \delta\right).$$

Using the same argument as in (12)–(14) we can derive that

$$P(E_2) < 2C_{\text{new},x}[\rho v(1-\delta^2)^{1/2}]^n,$$

and that

$$P(E_i) < i \cdot C_{\text{new},x}[\rho v(1-\delta^2)^{1/2}]^n \text{ for all } i = 1, \ldots, k. \quad (16)$$

Consider now events $B_1, \ldots B_k$ and evalutate $P(B_i)$, $i = 1, \ldots, k$:

$$P(B_i) = \int_{\mathbb{B}_n(v(1-\varepsilon),\boldsymbol{c})} p_{\text{new},x}(\boldsymbol{x})d\boldsymbol{x} \quad (17)$$
$$\leq \frac{C_{\text{new},x}\rho^n}{V_n(\mathbb{B}_n)} \int_{\mathbb{B}_n(v(1-\varepsilon),\boldsymbol{c})} d\boldsymbol{x}$$
$$= C_{\text{new},x}\rho^n \frac{V_n(\mathbb{B}_n(v(1-\varepsilon),\boldsymbol{c})}{V_n(\mathbb{B}_n)} = C_{\text{new},x}[\rho v(1-\varepsilon)]^n.$$

Recall that, for any sets $A_1, \ldots, A_d$, De Morgan's law states that:

$$\bigwedge_{i=1}^{d} A_i = \text{not } \left(\bigvee_{i=1}^{d} (\text{not } A_i)\right).$$

Therefore

$$P(A_1 \wedge A_2 \wedge \cdots \wedge A_d) =$$
$$1 - P((\text{not } A_1) \vee (\text{not } A_2) \vee \cdots (\text{not } A_d)).$$

Using the union bound rule (15), one can derive that

$$P(A_1 \wedge A_2 \wedge \cdots \wedge A_d) \geq 1 - \sum_{i=1}^{d} P(\text{not } A_i). \quad (18)$$

To complete the lemma, consider events

$$\text{not } E_1 : \left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_2\right)\right| \leq \delta$$
$$\text{not } E_2 : \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_3\right)\right| \leq \delta\right] \wedge \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_2}{\|\tilde{\boldsymbol{x}}_2\|}, \tilde{\boldsymbol{x}}_3\right)\right| \leq \delta\right]$$
$$\vdots \qquad \vdots$$
$$\text{not } E_{k-1} : \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_1}{\|\tilde{\boldsymbol{x}}_1\|}, \tilde{\boldsymbol{x}}_k\right)\right| \leq \delta\right] \wedge \cdots$$
$$\wedge \left[\left|\left(\frac{\tilde{\boldsymbol{x}}_{k-1}}{\|\tilde{\boldsymbol{x}}_{k-1}\|}, \tilde{\boldsymbol{x}}_k\right)\right| \leq \delta\right]$$
$$\text{not } B_1 : \|\tilde{\boldsymbol{x}}_1\| \geq (1-\varepsilon)v$$
$$\vdots \qquad \vdots$$
$$\text{not } B_k : \|\tilde{\boldsymbol{x}}_k\| \geq (1-\varepsilon)v$$

Given that $\boldsymbol{x}_i \in \mathbb{B}_n(v, \boldsymbol{c})$, we have that $\|\tilde{\boldsymbol{x}}_i\| \leq v$. It is hence clear that the event $[\text{not } E_1 \wedge \cdots \wedge \text{not } E_{k-1}]$ is contained in the event A_1 defined by (8) in the sense that any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{B}_n(v, \boldsymbol{c})$ for which

$$\text{not } E_1 \wedge \cdots \wedge \text{not } E_{k-1}$$

holds true must necessarily satisfy (8).

Therefore, according to (9), (18), we can write:

$$P(A_1) \geq P(\text{not } E_1 \wedge \cdots \wedge \text{not } E_{k-1}) \geq 1 - \sum_{i=1}^{k-1} P(E_i),$$

and

$$P(A_1 \wedge A_2) = P(A_1 \wedge \text{not } B_1 \wedge \cdots \text{not } B_k)$$
$$\geq P(\text{not } E_1 \wedge \cdots \wedge \text{not } E_{k-1} \wedge \text{not } B_1 \wedge \cdots \wedge \text{not } B_k \cdots)$$
$$\geq 1 - \sum_{i=1}^{k-1} P(E_i) - \sum_{i=1}^{k} P(B_i).$$

Substituting (16), (17) into the latter expressions one can now conclude that the lemma holds true. □

Remark 4: Lemma 1 reveals, in a general setting, the typicality of large "almost" or quasi- orthogonal bases in high-dimension (cf. [12]). Indeed, according to (11), if $\rho v(1-\varepsilon) < 1$, $\rho v(1-\delta^2)^{1/2} < 1$ then

$$|\cos(\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{x}_j - \boldsymbol{c})| = \frac{|(\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{x}_j - \boldsymbol{c})|}{\|\boldsymbol{x}_i - \boldsymbol{c}\|\|\boldsymbol{x}_j - \boldsymbol{c}\|}$$
$$\leq \frac{\delta v}{v^2(1-\varepsilon)^2} = \frac{\delta}{v(1-\varepsilon)^2}$$

with probability close to 1 if n is sufficiently large. Earlier works [17] (see also [18], [19]) showed that large ($k \gg n$) quasi-orthogonal bases exist. Here we follow our earlier results [12] and prove that almost orthogonal corteges of vectors whose cardinalily k grows exponentially with dimension n are *typical* in high dimension.

Our next result, Lemma 2 shows how this almost or quasi-orthogonality property can be used to estimate centroids of data clusters in high-dimensional datasets from few observations.

Lemma 2: Let $\mathcal{Y} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ be a set of k i.i.d. random vectors drawn from a distribution satisfying Assumption 2, and let $\delta, \varepsilon \in (0,1)$. Let

$$\bar{\boldsymbol{x}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{x}_i$$

be the empirical mean of the sample.

Then

$$P\left(L(k,\delta,\varepsilon) \leq \|\bar{\boldsymbol{x}} - \boldsymbol{c}\|^2 \leq U(k,\delta)\right) \geq 1 - R_{\varepsilon,\delta}(n,k,v,\rho,\delta,\varepsilon), \quad (19)$$

and

$$P\left(\|\bar{\boldsymbol{x}} - \boldsymbol{c}\|^2 \leq U(k,\delta)\right) \geq 1 - R_{\delta}(n,k,v,\rho,\delta), \quad (20)$$

where

$$U(k,\delta) = \frac{v^2}{k} + \frac{k-1}{k}v\delta$$
$$L(k,\delta,\varepsilon) = \frac{(1-\varepsilon)^2v^2}{k} - \frac{k-1}{k}v\delta$$

and

$$R_{\varepsilon,\delta}(n,k,v,\delta,\rho,\varepsilon) = C_{\text{new}} k[\rho v(1-\varepsilon)]^n + C_{\text{new}} \frac{k(k-1)}{2}\left[\rho v(1-\delta^2)^{1/2}\right]^n,$$

$$R_{\delta}(n,k,v,\rho,\delta) = C_{\text{new}} \frac{k(k-1)}{2}\left[\rho v(1-\delta^2)^{1/2}\right]^n.$$

Proof of Lemma 2. The Lemma is essentially contained in Lemma 1. Indeed, consider

$$\|\bar{\boldsymbol{x}} - \boldsymbol{c}\|^2 = (\bar{\boldsymbol{x}} - \boldsymbol{c}, \bar{\boldsymbol{x}} - \boldsymbol{c}) = \left(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{x}_i - \boldsymbol{c}, \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{x}_i - \boldsymbol{c}\right)$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\|\boldsymbol{x}_i - \boldsymbol{c}\|^2 + \frac{1}{k^2}\sum_{i\neq j}(\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{x}_j - \boldsymbol{c}).$$

According to Lemma 1 (statement (10)), the term

$$\left|\frac{1}{k^2}\sum_{i\neq j}(\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{x}_j - \boldsymbol{c})\right| \leq \frac{k-1}{k}v\delta$$

with probability $1 - R_\delta(n,k,v,\delta)$. This, together with the fact that $\|\boldsymbol{x}_i - \boldsymbol{c}\| \leq v$ for all $i = 1,\ldots,k$, prove (19). Similarly, statement (11) of Lemma 1 implies now that bound (20) holds true too. □

Theorem 2: [Learning from few examples] Consider a classifier F defined by (1), and let $\mathcal{U}_{\text{new}} = \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k\}$, $k \in \mathbb{N}$, $\boldsymbol{u}_i \in \mathcal{U}_{\text{new}}$, be a finite independent and identically distributed (i.i.d.) sample from a distribution P_{new}, and $\ell_{\text{new}} \in \mathcal{L}$ be a corresponding new label to be associated with the elements drawn from P_{new}.

Let $\mathcal{Y} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k\}$, $\boldsymbol{x}_i = f(\boldsymbol{u}_i)$, $i = 1,\ldots,k$ be a representation of the sample $\mathcal{U}_{\text{new}}$ in the classifier's latent space,

$$\bar{\boldsymbol{x}} = \frac{1}{k}\sum_i \boldsymbol{x}_i,$$

be the empirical mean of the representation, and let Assumption 2 hold. Finally, let $\delta \in (0,1)$ be a number satisfying

$$\Delta = \|\bar{\boldsymbol{x}}\| - \left(\frac{v^2}{k} + \frac{k-1}{k}v\delta\right)^{1/2} > 0.$$

Then the map

$$g^*:\ g^*(\boldsymbol{x}) = \begin{cases} \ell_{\text{new}}, & \text{if } \left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}\right) - \theta \geq 0 \\ g(\boldsymbol{x}), & \text{otherwise} \end{cases} \tag{21}$$

parameterised by

$$\theta \in [\max\{\Delta - v, 0, \}, \Delta]$$

is a solution of Problem 2 with

$$p_n = \left(1 - \frac{C_{\text{new},x}}{2}\left[\rho\left(v^2 - (\Delta-\theta)^2\right)^{1/2}\right]^n\right) \times \left(1 - C_{\text{new}}\frac{k(k-1)}{2}\left[\rho v(1-\delta^2)^{1/2}\right]^n\right), \tag{22}$$

$$p_e = 1 - \frac{C_x}{2}\left[r\left(1-\theta^2\right)^{1/2}\right]^n.$$

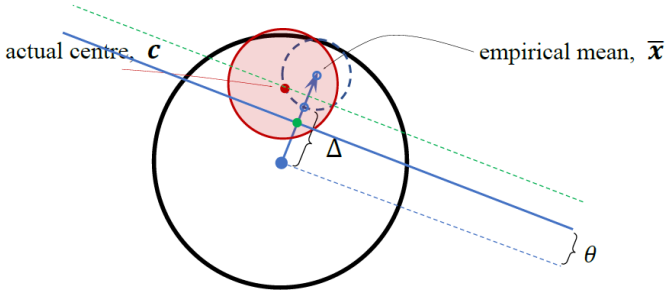

Fig. 2. Illustration to the proof of Theorem 2. Pink filled circle shows $\mathbb{B}_n(v,\boldsymbol{c})$, blue dashed disc shows the domain where the true centre $\boldsymbol{c}$ is located (with high probability), and blue solid line shows the hyperplane $\left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}\right) - \theta = 0$.

Proof of Theorem 2. According to Lemma 2, the probability that the centre $\boldsymbol{c}$ is within $\left(\frac{v^2}{k} + \frac{k-1}{k}v\delta\right)^{1/2}$ from the empirical mean $\bar{\boldsymbol{x}}$ is at least

$$\left(1 - C_{\text{new}}\frac{k(k-1)}{2}\left[\rho v(1-\delta^2)^{1/2}\right]^n\right).$$

Suppose that the above event occurs. This implies that the hyperplane

$$\boldsymbol{x}:\ \left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}\right) - \theta = 0$$

is at least $\Delta - \theta$ away from the hyperplane with the same normal, $\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}$, and which is passing through the centre $\boldsymbol{c}$ of the ball $\mathbb{B}_n(v,\boldsymbol{c})$ (see Fig. 2).

The probability that an element drawn from the distribution P_{new} would have a representation $\boldsymbol{x}$ for which $\left(\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|}, \boldsymbol{x}\right) - \theta > 0$ is hence at least

$$1 - \frac{C_{\text{new},x}}{2}\left[\rho\left(v^2 - (\Delta-\theta)^2\right)^{1/2}\right]^n.$$

This justifies the expression for p_n in (22).

Similarly, the probability that an element drawn from the distribution P_u would be assigned a label ℓ_{new} is at most

$$\frac{C_x}{2}\left[r\left(1-\theta^2\right)^{1/2}\right]^n.$$

Hence the expression for p_e follows. □

Remark 5: According to Theorem 2 and similar to the case covered by Theorem 1, under appropriate and reasonable assumptions, the probabilities of success in the task of learning from few examples approach 1 exponentially fast as n grows.

C. Discussion

Having introduced our main theoretical results, let us now briefly relate these results to existing literature on few-shot learning and outline future potential direction s.

1) Matching and prototypical networks: Theorems 1, 2 and few-shot learning algorithms (6), (21), which these theorems relate to, show striking similarity to approaches presented and empirically studied in [6], [7]. In the case of one-shot learning [6], Theorem 1 with $k = 1$ applies, whereas in the case of few-shot learning, [7], Theorem 2 could be more appropriate for explaining and interpreting why few-shot learning works.

2) Object models and the challenge of generalisation: Our results show that significant understanding and insights into why and when large-scale and highly expressive AI systems, including deep neural networks, can generalise well from just few examples can be gained if some loose assumptions are introduced on the data models. In our case, these assumptions, are that 1) the probability distributions of objects' representations in the system's latent space are supported on some balls (or ellipsoids, subject to a coordinate transformation), and 2) these probability distributions are not degenerate in the sense of Assumptions 1, 2. Going forward, one can consider further straightforward generalisations in which the objects are modeled by mixtures of these models. These generalisations, are however, beyond the scope of the current work.

In addition, our current work, by focusing on what can and what cannot be learned from few examples in randomised settings, provides insights into why stochastic configuration networks may be so successful in practice [20], [21]: practically relevant functions we are interested to learn may have a "compact" structure, and the process of stochastic configuration could be viewed as an efficient mechanism that is capable to learn this structure from data step-by-step.

3) Learning to learn: In addition to explaining why few-shot learning models work and why large-scale deep learning models may generalise so well, our present work *presents high-level training requirements* for a model that is trained to learn from few examples. These requirements are specified in Assumptions 1 and 2. If a network is trained so that object representations in its latent space satisfy Assumptions 1 and 2 with appropriate relevant constants then Theorems 1, 2 guarantee that such models can indeed learn from mere few or single examples. Importantly, training of networks to satisfy Assumptions 1 and 2 can be posed within the standard empirical risk minimisation framework. A very similar approach has been pursued in [7], [6], albeit heuristically.

IV. Conclusion

This work presents a formal treatment of the challenges of few-shot and one-shot learning and generalisation in large-scale modern AI models. We provided formal statements of these learning problems and showed that high dimensionality and geometry of objects' representations in the systems' latent spaces along with some non-degeneracy conditions are key determinants explaining when and why such learning is possible.

Our results suggest that neural networks' generalisation capabilities are intrinsically linked with internal regularities in the data sets and also with representations of these regularities in the networks' latent spaces. The results reveal an important characteristic of this important regularity: if an object has a "compact" representation in the network's latent space then such object can be learned from just few or even single example. Absence of such compact representations may require exponentially large training samples to learn from.

References

[1] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[2] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, X. Xianghua, M. Jones, and K. Gary, Eds. BMVA Press, September 2015, pp. 41.1–41.12. [Online]. Available: https://dx.doi.org/10.5244/C.29.41

[3] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

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

[5] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 033–30 038, 2020.

[6] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[8] A. Gorban and I. Tyukin, "Stochastic separation theorems," *Neural Networks*, vol. 94, pp. 255–259, 2017.

[9] B. Grechuk, A. Gorban, and I. Tyukin, "General stochastic separation theorems with optimal bounds," *Neural Networks*, vol. 138, pp. 33–56, 2021. [Online]. Available: https://doi.org/10.1016/j.neunet.2021.01.034

[10] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American mathematical society*, vol. 39, no. 1, pp. 1–49, 2002.

[11] I. Y. Tyukin, A. N. Gorban, C. Calvo, J. Makarova, and V. A. Makarov, "High-dimensional brain. A tool for encoding and rapid learning of memories by single neurons," *Bulletin of Mathematical Biology*, vol. 81, pp. 4856–4888, 2019. [Online]. Available: https://doi.org/10.1007/s11538-018-0415-5

[12] A. Gorban, I. Tyukin, D. Prokhorov, and K. Sofeikov, "Approximation with random bases: Pro et contra," *Information Sciences*, vol. 364–365, pp. 129–145, 2016.

[13] I. Y. Tyukin, A. N. Gorban, K. Sofeikov, and I. Romanenko, "Knowledge transfer between artificial intelligence systems," *Frontiers of Neurorobotics*, vol. 12, Article 49, 2018.

[14] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of AI systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.

[15] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, "The unreasonable effectiveness of small neural ensembles in high-dimensional brain," *Physics of Life Reviews*, 2018.

[16] A. N. Gorban, V. Makarov, and I. Tyukin, "High-dimensional brain in a high-dimensional world: Blessing of dimensionality," *Entropy*, vol. 22, no. 1, p. 82, 2020.

[17] P. Kainen and V. Kurkova, "Quasiorthogonal dimension of euclidian spaces," *Appl. Math. Lett.*, vol. 6, no. 3, pp. 7–10, 1993.

[18] P. C. Kainen and V. Kurkova, "Quasiorthogonal dimension," in *Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy etc. Methods and Their Applications*. Springer, 2020, pp. 615–629.

[19] P. C. Kainen, "Utilizing geometric anomalies of high dimension: When complexity makes computation easier," in *Computer Intensive Methods in Control and Signal Processing*. Springer, 1997, pp. 283–294.

[20] D. Wang and M. Li, "Stochastic configuration networks: Fundamentals and algorithms," *IEEE Transactions on Cybernetics*, vol. 47, no. 10, pp. 3466–3479, 2017.

[21] C. Huang, Q. Huang, and D. Wang, "Stochastic configuration networks based adaptive storage replica management for power big data processing," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 373–383, 2019.