March 10, 2020; Pages:1–29

Approximate is Good Enough: Probabilistic Variants of Dimensional and Margin Complexity

Pritish Kamath PRITISH@TTIC.EDU
Omar Montasser OMAR@TTIC.EDU
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago, 6045 S Kenwood Ave, Chicago, IL 60637

Abstract

We present and study approximate notions of dimensional and margin complexity, which correspond to the minimal dimension or norm of an embedding required to *approximate*, rather then exactly represent, a given hypothesis class. We show that such notions are not only sufficient for learning using linear predictors or a kernel, but unlike the exact variants, are also necessary. Thus they are better suited for discussing limitations of linear or kernel methods.

Keywords: Kernel Methods, Dimensional Complexity, Margin Complexity, Random Features

1. Introduction

A possible approach to learning is to choose some feature map $\varphi(x)$, or equivalently some kernel $K(x, x') := \langle \varphi(x), \varphi(x') \rangle$, appropriate for the problem, and then reduce the problem of learning, to that of learning a linear predictor, or a low (Euclidean or Hilbert) norm linear predictor, with respect to this embedding. Such an approach is often successful in practice, and is the basis of "kernel methods". But what are the inherent limits of such an approach? Are there easily learnable hypothesis classes that *cannot* be learnt using such an approach, or perhaps require many more samples for learning, no matter what feature map or kernel is used? This classic question about the limits of kernel methods has been explored by, e.g. Ben-David et al. (2002), and has lead to the notions of dimensional and margin complexity of a hypothesis class— these correspond to the minimal dimension and minimal norm (respectively) of a feature space sufficient to *exactly* represent all hypotheses in the class as linear predictors (see precise definitions in Section 2). Dimensional and margin complexity have also been studied in communication complexity (See e.g., Forster and Simon, 2006; Forster et al., 2003; Sherstov, 2008; Razborov and Sherstov, 2010). Questions about the limits of kernel methods have resurfaced in recent years, in the context of understanding the advantage of deep learning over kernel methods, and identifying hypothesis classes that are learnable by training a neural network (using an efficient and simple training procedure) but that are not learnable, or at least not without many more samples, using any kernel or feature map (Allen-Zhu and Li, 2019, 2020; Yehudai and Shamir, 2019).

While the standard notions of dimensional and margin complexity are *sufficient* for learning by reduction to linear learning, they might not be *necessary* for such an approach. This is because these notions insist on a feature map that can be used to *exactly* represent all hypotheses in the class, without any errors. But for learning, it is sufficient to only *approximate* the hypotheses, up to a small error ε. Furthermore, once we allow small errors, we might want to consider *randomized* rather than *deterministic* feature maps or kernels. This is not only a hypothetical possibility—examples of specific randomized feature maps and kernels include Random Fourier Features (Rahimi and Recht, 2008), the Conjugate Kernel (Daniely, 2017), and the Neural Tangent Kernel at a random initialized

neural network (Jacot et al., 2018). One might ask if such randomized approximate embedding are in fact more powerful, or whether perhaps they can always be de-randomized and made exact. In this paper we establish (Theorem 6, combined with Theorem 11) that randomized approximate embedding are indeed more powerful: we show that learning *is* possible using a randomized feature map, even for a hypothesis class for which no *exact* low dimensional representation exists (i.e. with a very high, or even infinite, dimensional complexity). In order to truly understand the power of kernel methods and reduction to linear learning, we must therefore also allow for such randomized feature maps and kernels, and understand their power and limitations.

In this paper we propose and study relaxed notions of dimensional and margin complexity that (a) allow for randomized feature maps; and (b) can be shown to be not only sufficient, but also necessary for learning by reduction to linear or kernel methods, and so yield strong lower bounds on the power of such an approach. In discussing approximation of a hypothesis class, we must consider the loss used, and we study both classification problems with respect to a hard (0/1) loss, as well as classification and regression with continuous losses such as the hinge and squared loss.

In order to be able to discuss a necessary condition for "learning by reduction to linear or kernel methods" we must precisely define what we mean by this phrase. We do so in Section 3. We consider both distribution-dependent and distribution-independent learning. Correspondingly, we define both distribution-dependent and distribution-independent approximate dimensional and margin complexity (in Section 2). Our complexity definitions are justified by showing how they are both necessary and sufficient (in a sense) for learning by reduction to kernel or linear methods. We also show how the distribution-dependent approximate dimension complexity lower bounds linear and kernel learning in a very broad sense, and with respect to a generic loss function. In Section 4 we further show how this complexity measure can be lower bounded, in turn, by other well studied complexity measures, providing for a generic way of obtaining strong lower bounds on the power of kernel methods.

Our generic lower bound approach mirrors, to a large extent, the lower bound on the sample complexity of kernel based learning in several recent papers exploring the power of deep learning versus kernel method (Allen-Zhu and Li, 2019, 2020; Yehudai and Shamir, 2019). We distil the approach to a crisp complexity measure, which simplifies making such lower bound claims on specific hypothesis classes, and can also lead to stronger statements—we demonstrate this by strengthening the lower bound and resolving an open question of Yehudai and Shamir (2019). Our lower bound is stated in terms of the Statistical Query dimension, as defined by Blum et al. (1994), making a concrete connection between these complexity measures ("dimensionalities"). Our treatment also highlights a potential deficiency of this approach: although we can establish lower bounds for learning w.r.t. the squared loss, using the same technique to establish a strong lower bound on learning w.r.t. the 0/1 loss would resolve a long-standing question in circuit complexity theory and thus seems much more difficult.

We emphasize that when we speak of "linear learning" we refer to learning by minimizing the loss over all linear predictors without any regularization, and when we refer to "kernel learning" or "norm based learning" we are specifically referring to constraining or regularizing the Euclidean or Hilbert norm of linear predictors. Learning using regularized linear predictors with other regularizers can be much more powerful—e.g. any (finite) hypothesis class can be optimally learned using ℓ_1 regularized learning with a feature map with dimension corresponding to the cardinality of the hypothesis class. But this is not much different than using the hypothesis class itself, and we cannot use the "kernel trick" in order to avoid an explicit representation and search over this

very high dimensional feature space. In this paper, we are only concerned with (low dimensional) unregularized and ℓ_2 regularized (kernel based) learning.

Throughout the paper, we are not overly concerned with the precise dependence on the "error parameter" ε. Although we always explicitly note the dependence on ε, we think of it as a small constant, perhaps 0.01, and do not worry about factors which are polynomial in ε. In this paper, we only refer to learning and approximating *in expectation*—it is possible to define and relate approximating and learning *with high probability* instead, but we avoid doing so for notational simplicity.

Notations. We refer to hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ over a domain $\mathcal{X}$ and label set $\mathcal{Y}$. When $\mathcal{Y}$ is $\mathbb{R}$ or $\{1,-1\}$, and $|\mathcal{X}|$ and $|\mathcal{H}|$ are finite, we associate $\mathcal{H}$ with a matrix $M_{\mathcal{H}} \in \mathbb{R}^{\mathcal{H}\times\mathcal{X}}$ defined as $M_{\mathcal{H}}(h,x) := h(x)$. We consider *loss* functions of the form $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. In particular, we consider the 0/1 loss $\ell_{0\text{-}1}(\widehat{y}, y) := \mathbb{1}\{\widehat{y}y \leq 0\}$, margin loss $\ell_{\mathrm{mgn}}(\widehat{y}, y) := \mathbb{1}\{\widehat{y}y \leq 1\}$ and hinge loss $\ell_{\mathrm{hinge}}(\widehat{y}, y) := \max\{0, 1-\widehat{y}y\}$ for binary labels $\mathcal{Y} = \{1,-1\}$, and the squared loss $\ell_{\mathrm{sq}}(\widehat{y}, y) := \frac{1}{2}(\widehat{y}-y)^2$, for $\mathcal{Y} \subseteq \mathbb{R}$. A loss ℓ is said to be L-Lipschitz if $|\ell(a,y)-\ell(a',y)| \leq L|a-a'|$ for all $a, a' \in \mathbb{R}$ and $y \in \mathcal{Y}$.

We view learning algorithms as operating on a set of samples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ drawn i.i.d. from a distribution $\mathscr{D}$ over $\mathcal{X} \times \mathcal{Y}$. We say that $\mathscr{D}$ is *realizable* w.r.t. a hypothesis class $\mathcal{H}$, to mean that $(x,y) \sim \mathscr{D}$ is sampled by first sampling $x \sim \mathcal{D}$ (for some $\mathcal{D}$) and setting $y = h_*(x_i)$ for some $h_* \in \mathcal{H}$. We always use $\mathscr{D}$ to denote a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{D}$ to denote its marginal over $\mathcal{X}$. The *population loss* of a predictor $g : \mathcal{X} \to \mathbb{R}$ w.r.t. a loss ℓ is $\mathcal{L}^{\ell}_{\mathscr{D}}(g) := \mathbb{E}_{(x,y)\sim\mathscr{D}}\ \ell(g(x), y)$ whereas its *empirical loss* is $\mathcal{L}^{\ell}_{S}(g) := \frac{1}{|S|}\sum_{(x,y)\in S}\ \ell(g(x), y)$. If $\mathscr{D}$ is realizable and sampled as $(x, h(x))$ with $x \sim \mathcal{D}$ and $h \in \mathcal{H}$, we define an alternate notation for $\mathcal{L}^{\ell}_{\mathscr{D}}(g)$ as $\mathcal{L}^{\ell}_{\mathcal{D},h}(g) := \mathbb{E}_{x\sim\mathcal{D}}\left[\ell(g(x), h(x))\right]$.

2. Dimension & Margin Complexities and their Probabilistic Variants

We recall the definitions of the dimension and margin complexities of a hypothesis class and introduce their probabilistic variants. Our definitions of the error-free notions are also stated in terms of a loss function so that we can then extend them to allow errors.

2.1. Dimension Complexity

Definition 1 *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and a loss ℓ. The* dimension complexity $\mathsf{dc}^{\ell}(\mathcal{H})$ *is the smallest d for which there exists an embedding $\varphi : \mathcal{X} \to \mathbb{R}^d$ and a map $w : \mathcal{H} \to \mathbb{R}^d$ such that for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$, it holds that $\ell(\langle w(h), \varphi(x)\rangle, h(x)) = 0$.*

For classification problems ($\mathcal{Y} = \{1,-1\}$) our definition coincides with the standard definition of dimensional complexity (equivalent to $\mathsf{sign\text{-}rank}(M_{\mathcal{H}})$) for $\ell = \ell_{0\text{-}1}$, and we will denote $\mathsf{dc}(\mathcal{H}) := \mathsf{dc}^{\ell_{0\text{-}1}}(\mathcal{H})$. For finite hypothesis classes we also have $\mathsf{dc}(\mathcal{H}) = \mathsf{dc}^{\ell_{\mathrm{mgn}}}(\mathcal{H}) = \mathsf{dc}^{\ell_{\mathrm{hinge}}}(\mathcal{H})$. For regression problems ($\mathcal{Y} = \mathbb{R}$), e.g. with the ℓ_{sq} loss, $\mathsf{dc}^{\ell_{\mathrm{sq}}}(\mathcal{H})$ coincides with $\mathsf{rank}(M_{\mathcal{H}})$.

Definition 2 *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, a loss ℓ and a parameter $\varepsilon \geq 0$.*

Probabilistic Distributional Dimension Complexity. $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon}(\mathcal{H})$*, parameterized by a distribution $\mathcal{D}$ over $\mathcal{X}$, is the smallest d for which there exists a distribution $\mathcal{P}$ over embeddings $\varphi : \mathcal{X} \to \mathbb{R}^d$ such that for all $h \in \mathcal{H}$,*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathbb{R}^d}\ \mathcal{L}^{\ell}_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle)\right] \leq \varepsilon\,. \tag{1}$$

Probabilistic Dimension Complexity. $\mathsf{dc}_\varepsilon^\ell(\mathcal{H})$ *is the smallest* d *for which there exists a distribution* $\mathcal{P}$ *over embeddings* $\varphi : \mathcal{X} \to \mathbb{R}^d$ *such that for all distributions* $\mathcal{D}$ *over* $\mathcal{X}$ *and all* $h \in \mathcal{H}$*, Equation (1) above holds.*

Again, for classification $\mathcal{Y} = \{1, -1\}$ we denote $\mathsf{dc}_\varepsilon(\mathcal{H}) = \mathsf{dc}_\varepsilon^{\ell_{0\text{-}1}}(\mathcal{H})$ and $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}) = \mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{0\text{-}1}}(\mathcal{H})$, and at least for finite hypothesis classes these also agree with the complexities with respect to losses ℓ_{mgn} and ℓ_{hinge}. Note that $\mathsf{dc}_\varepsilon^\ell(\mathcal{H})$ is different from simply $\sup_{\mathcal{D}} \mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$. In particular, note the difference in order of quantifiers.

$$\mathsf{dc}_\varepsilon^\ell(\mathcal{H}) \;:=\; \min\; d \qquad \exists \mathcal{P} \quad \forall \mathcal{D} \quad \forall h \quad \exists w | \varphi, h$$

$$\sup_{\mathcal{D}} \mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H}) \;:=\; \min\; d \qquad \forall \mathcal{D} \quad \exists \mathcal{P} \quad \forall h \quad \exists w | \varphi, h$$

2.2. Margin Complexity

Margin complexity is defined in terms of embeddings $\varphi : \mathcal{X} \to \mathbb{H}$, for any Hilbert space $\mathbb{H}$, thereby also allowing infinite dimensional embeddings, typically represented via a kernel $K_\varphi(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathbb{H}}$. The sup-norm of the embedding is defined as $\|\varphi\|_\infty := \sup_{x \in \mathcal{X}} \|\varphi(x)\|_{\mathbb{H}} = \sup_{x \in \mathcal{X}} \sqrt{K_\varphi(x, x)}$. For a parameter $R \in \mathbb{R}_{\geq 0}$, let $\mathcal{B}(\mathbb{H}; R) := \{w \in \mathbb{H} : \|w\|_{\mathbb{H}} \leq R\}$ be a norm ball of radius R in the Hilbert space.

Definition 3 *Fix a hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *and a loss* ℓ*. The* margin complexity $\mathsf{mc}^\ell(\mathcal{H})$ *is the smallest* R *for which there exists an embedding* $\varphi : \mathcal{X} \to \mathbb{H}$ *and a map* $w : \mathcal{H} \to \mathbb{H}$ *with* $\|\varphi\|_\infty \leq 1$ *and* $\|w\|_\infty \leq R$ *such that for all* $h \in \mathcal{H}$ *and* $x \in \mathcal{X}$*, it holds that* $\ell(\langle w(h), \varphi(x) \rangle, h(x)) = 0$*.*

This definition does not make sense for the $\ell_{0\text{-}1}$ loss, since $\ell_{0\text{-}1}$ is scale-invariant. However, in the case of $\mathcal{Y} = \{1, -1\}$, it coincides with the standard definition of margin complexity for the margin loss ℓ_{mgn} (and hinge loss ℓ_{hinge}), and we denote $\mathsf{mc}(\mathcal{H}) := \mathsf{mc}^{\ell_{\mathrm{mgn}}}(\mathcal{H})$. For the squared loss ℓ_{sq}, the definition coincides with the $\gamma_{2:\ell_1 \to \ell_\infty}$ norm (Jameson, 1987), a.k.a. the "max norm" (Srebro and Shraibman, 2005). Especially with a general loss function, "mc" is really a form of "norm-complexity", but we still refer to it as "margin complexity" and use mc since it does capture the (inverse) margin when $\ell = \ell_{\mathrm{mgn}}$ and this term is already widely used in the literature.

Definition 4 *Fix a hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*, a loss* ℓ *and a parameter* $\varepsilon \geq 0$*.*

Probabilistic Distributional Margin Complexity. $\mathsf{mc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$*, parameterized by a distribution* $\mathcal{D}$ *over* $\mathcal{X}$*, is the smallest* R *for which there exists a distribution* $\mathcal{P}$ *over embeddings* $\varphi : \mathcal{X} \to \mathbb{H}$ *with* $\|\varphi\|_\infty \leq 1$ *such that for all* $h \in \mathcal{H}$*,*

$$\mathop{\mathbb{E}}_{\varphi \sim \mathcal{P}} \left[\inf_{w \in \mathcal{B}(\mathbb{H};R)} \mathcal{L}^\ell_{\mathcal{D},h}(\langle w, \varphi(\cdot) \rangle) \right] \leq \varepsilon \,. \tag{2}$$

Probabilistic Margin Complexity. $\mathsf{mc}_\varepsilon^\ell(\mathcal{H})$ *is the smallest* R *for which there exists a distribution* $\mathcal{P}$ *over embeddings* $\varphi : \mathcal{X} \to \mathbb{H}$ *with* $\|\varphi\|_\infty \leq 1$ *such that for all distributions* $\mathcal{D}$ *over* $\mathcal{X}$ *and all* $h \in \mathcal{H}$*, Equation (2) above holds.*

When $\mathcal{Y} = \{1, -1\}$, we denote $\mathsf{mc}_\varepsilon(\mathcal{H}) = \mathsf{mc}_\varepsilon^{\ell_{\mathrm{mgn}}}(\mathcal{H})$ and $\mathsf{mc}_\varepsilon^{\mathcal{D}}(\mathcal{H}) = \mathsf{mc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{mgn}}}(\mathcal{H})$.

2.3. Relationship between Probabilistic Dimension & Margin Complexity

A classic result attributed to Arriaga and Vempala (1999) and Ben-David et al. (2002) shows that

$$\mathsf{dc}(\mathcal{H}) \leq \mathsf{mc}(\mathcal{H})^2 \cdot \mathcal{O}(\log|\mathcal{H}||\mathcal{X}|)\,. \tag{3}$$

This result is proved by an application of the lemma of Johnson and Lindenstrauss (1984). The term of $\mathcal{O}(\log|\mathcal{H}||\mathcal{X}|)$ comes up due to a union bound over all pairs of $(x,h) \in \mathcal{X}\times\mathcal{H}$. Although the result can be seen as establishing a tight connection between the dimension and margin complexity, it is not applicable with continuous (or simply infinite) domains, and we are not aware of any way of avoiding this dependence on the cardinality of the domain.

As a first application of our probabilistic notions, we show how this bypasses the cardinality dependence when allowing a randomized feature map.

Lemma 5 (Relating dc and mc) *For all $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and parameters $\varepsilon, \eta > 0$,*

$$\begin{aligned}
&\textit{(i)}\ \mathsf{dc}_{\varepsilon+\eta}(\mathcal{H}) &&\leq\ \mathsf{mc}_{\varepsilon}(\mathcal{H})^2 \cdot \mathcal{O}\left(\log(1/\eta)\right),\\
&\textit{(ii)}\ \mathsf{dc}^{\ell}_{\varepsilon+\eta}(\mathcal{H}) &&\leq\ \mathsf{mc}^{\ell}_{\varepsilon}(\mathcal{H})^2 \cdot \mathcal{O}\left(L/\eta\right)^2 \quad \textit{for any } L\textit{-Lipschitz loss } \ell, \textit{ and}\\
&\textit{(iii)}\ \mathsf{dc}^{\ell_{\mathrm{sq}}}_{\varepsilon+\eta}(\mathcal{H}) &&\leq\ \mathsf{mc}^{\ell_{\mathrm{sq}}}_{\varepsilon}(\mathcal{H})^2 \cdot \mathcal{O}\left((\varepsilon+\eta)/\eta^2\right).
\end{aligned}$$

Analogous statements relating $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon+\eta}$ and $\mathsf{mc}^{\mathcal{D},\ell}_{\varepsilon}$ hold as well for any distribution $\mathcal{D}$ over $\mathcal{X}$.

The proof is similar to that of Ben-David et al. (2002) in its use of the lemma of Johnson and Lindenstrauss (1984). We defer the proof details to Appendix A. The random feature map used here is analogous to random features used in practice to approximate kernels (Rahimi and Recht, 2007).

2.4. Separations between Deterministic and Probabilistic Dimension Complexity

We show that the probabilistic variants $\mathsf{dc}_{\varepsilon}$ and $\mathsf{dc}^{\mathcal{D}}_{\varepsilon}$ can sometimes be significantly smaller than the classic notion of dc. We show that dimension complexity can be exponentially larger than probabilistic dimension complexity (with respect to $\ell_{0\text{-}1}$). Moreover, if we focus on the distributional version, then in fact dimension complexity can be "infinitely larger" than probabilistic distributional dimension complexity and moreover this separation holds for different losses such as $\ell_{0\text{-}1}$, ℓ_{sq} and ℓ_{hinge}, as well as for margin complexity.

Theorem 6 (Exponential Distribution Independent Gap) *For $\mathcal{X} = \{1,-1\}^n$, there exists a hypothesis class $\mathcal{H} \subseteq \{1,-1\}^{\mathcal{X}}$ with $|\mathcal{H}| = 2^n$ such that, for all $\varepsilon \in (0, 1/2)$,*

$$\mathsf{dc}_{\varepsilon}(\mathcal{H}) \ \leq\ \mathcal{O}\left(n^4/\varepsilon\right) \qquad \text{and} \qquad \mathsf{dc}(\mathcal{H}) \ \geq\ 2^{\Omega(n^{1/4})}.$$

Theorem 7 ("Infinite" Distribution Dependent Gap) *For every n, there exist hypothesis classes $\mathcal{H} \subseteq \{1,-1\}^{\mathcal{X}}$ with $|\mathcal{H}| = |\mathcal{X}| = 2^n$ such that for all $\varepsilon \in (0, 1/2)$,*

$$\begin{aligned}
&\sup_{\mathcal{D}}\ \mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon}(\mathcal{H}) \ \leq\ \mathcal{O}\left(1/\varepsilon^2\right) \quad \text{and} \quad \mathsf{dc}^{\ell}(\mathcal{H}) \ \geq\ 2^{\Omega(n)} \quad \textit{for } \ell \in \{\ell_{0\text{-}1}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}\\
&\sup_{\mathcal{D}}\ \mathsf{mc}^{\mathcal{D},\ell}_{\varepsilon}(\mathcal{H}) \ \leq\ \mathcal{O}\left(1/\varepsilon^2\right) \quad \text{and} \quad \mathsf{mc}^{\ell}(\mathcal{H}) \ \geq\ 2^{\Omega(n)} \quad \textit{for } \ell \in \{\ell_{\mathrm{mgn}}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}
\end{aligned}$$

We prove Theorem 6 as follows (full details in Appendix B.1): We define another notion of probabilistic dimension complexity that has a stronger requirement of pointwise correctness and hence is larger than dc_ε. This notion is equivalent to *probabilistic sign-rank* studied in communication complexity. In particular, Alman and Williams (2017) showed that if the function $E_{\mathcal{H}} : \mathcal{H} \times \mathcal{X} \to \{1, -1\}$ defined as $E_{\mathcal{H}}(h, x) := h(x)$ is computable by a "small" depth-2 threshold circuit (for some encoding of $\mathcal{H}$ and $\mathcal{X}$ into bits), then $M_{\mathcal{H}}$ has "small" probabilistic sign-rank. The theorem follows from a lower bound on sign-rank shown by Chattopadhyay and Mande (2018) for matrices that are computable by "small" depth-2 threshold circuits. The hypothesis class $\mathcal{H}$ witnessing this separation is a class of *decision lists of conjunctions over disjoint variables*.

We prove Theorem 7 as follows (full details in Appendix B.2): We use the "covering lemma" of Haussler (1995) to show that the probabilistic distributional dimension complexity of any class can be bounded, albeit exponentially, in terms of the VC dimension, establishing the following Lemma:

Lemma 8 ($\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}$ and $\mathsf{mc}_\varepsilon^{\mathcal{D},\ell}$ versus VC-dim) *There exists universal constants c, K such that for all hypothesis classes $\mathcal{H} \subseteq \{1, -1\}^{\mathcal{X}}$, parameter $\varepsilon > 0$ and all losses $\ell \in \{\ell_{0\text{-}1}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}$ (in case of dc) and $\ell \in \{\ell_{\mathrm{mgn}}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}$ (in case of mc),*

$$\sup_{\mathcal{D}} \mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H}) \;,\; \sup_{\mathcal{D}} \mathsf{mc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H}) \;\leq\; c \cdot \mathsf{VC\text{-}dim}(\mathcal{H}) \left(\frac{K}{\varepsilon}\right)^{\mathsf{VC\text{-}dim}(\mathcal{H})} .$$

This is in contrast to the exact dimensional complexity, which can be polynomially large in $|\mathcal{H}||\mathcal{X}|$ even for classes of bounded VC dimension Alon et al. (2016). Theorem 7 now follows by considering a hypothesis class with VC-dimension 2 with dimensional complexity of $2^{\Omega(n)}$.

The construction in Theorem 6 uses extremely large magnitude features and weights, whereas the construction in Theorem 7 uses bounded magnitude of features and weights, but relies on having a known marginal $\mathcal{D}$ over $\mathcal{X}$. Our theorems therefore leave open the following questions.

Open Questions. Is there an "infinite" separation between distribution independent dc_ε and exact dc? Is there a large (even finite) separation between distribution independent mc_ε and exact mc? Also between distribution independent $\mathsf{dc}_\varepsilon^\ell$ and exact dc^ℓ for $\ell \in \{\ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}$? Can the distribution independent dc_ε also be bounded in terms of the VC dimension?

3. Linear & Kernel Learnability with Probabilistic Embeddings

We now turn to precisely defining learning by reduction to Linear Learning or Kernel Learning. These notions serve as the primary motivation for our work, and their definitions guided the definitions of the other complexity notions we consider.

3.1. Linear Learning Complexity

Linear learning with a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ boils down to relying on a learning rule of the form

$$\textsc{Erm}_\varphi^\ell(S) \;:=\; \operatorname{argmin}_{w \in \mathbb{R}^d} \;\; \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle)\,, \tag{4}$$

where we require generalization for *any* minimizer of the empirical error. We formalize the *Linear Learning Complexity* of a hypothesis class $\mathcal{H}$ as the minimal sample complexity of *any* learning rule of the form (4).

Definition 9 *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, a loss ℓ and parameter $\varepsilon > 0$.*

Distributional Linear Learning Complexity $\mathsf{Lin}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$*, parameterized by a distribution* $\mathcal{D}$ *over* $\mathcal{X}$*, is the smallest* m *for which there exists a distribution* $\mathcal{P}$ *over embeddings* $\varphi : \mathcal{X} \to \mathbb{R}^d$ *(for some* $d \in \mathbb{N}$*) such that for all realizable distributions* $\mathscr{D}$ *over* $\mathcal{X} \times \mathcal{Y}$ *with marginal* $\mathcal{D}$ *over* $\mathcal{X}$*,*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\sup_{w\in \textsc{Erm}^\ell_\varphi(S)} \mathcal{L}^\ell_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle) \right] \le \varepsilon. \tag{5}$$

Linear Learning Complexity $\mathsf{Lin}_\varepsilon^{\ell}(\mathcal{H})$ *is the smallest* m *for which there exists a distribution* $\mathcal{P}$ *over embeddings* $\varphi : \mathcal{X} \to \mathbb{R}^d$ *(for some* $d \in \mathbb{N}$*) such that for all realizable distributions* $\mathscr{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*, Equation (5) above holds.*

For $\mathcal{Y} = \{1, -1\}$*, we denote* $\mathsf{Lin}_\varepsilon(\mathcal{H}) = \mathsf{Lin}_\varepsilon^{\ell_{0\text{-}1}}(\mathcal{H})$ *and* $\mathsf{Lin}_\varepsilon^{\mathcal{D}}(\mathcal{H}) = \mathsf{Lin}_\varepsilon^{\mathcal{D},\ell_{0\text{-}1}}(\mathcal{H})$.

To see more explicitly how low dimensional complexity is sufficient for linear learning, we also consider a stronger definition which requires that learning can be ensured by relying on linear dimension based generalization guarantees. Recall that for a bounded or Lipschitz loss ℓ we have that for any distribution $\mathscr{D}$ (c.f. Shalev-Shwartz and Ben-David, 2014),

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\sup_{w\in\mathbb{R}^d} \left(\mathcal{L}^\ell_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle) - \mathcal{L}^\ell_S(\langle w, \varphi(\cdot)\rangle) \right) \right] \le C^\ell_{\mathsf{dc}} \sqrt{\frac{d}{m}} \tag{6}$$

for some constant C^ℓ_{dc} that depends on either the range or Lipschitz constant of the loss. We note that the square-root dependence in the right-hand side can be improved to a nearly linear dependence when the empirical error is small, as it would be in our realizable setting. This would yield a better polynomial dependence on the error parameter ϵ. Since we are less concerned here with the precise polynomial dependence on the error parameter, we refer only to the simpler uniform bound (6).

Definition 10 *Fix a hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$*, a loss* ℓ *that is either bounded or Lipschitz over the domain, and parameter* $\varepsilon > 0$*. The* **Guaranteed Linear Learning Complexity** $\mathsf{gLin}_\varepsilon^{\ell}(\mathcal{H})$ *and* **Distributional Guaranteed Linear Learning Complexity** $\mathsf{gLin}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ *are defined as in Definition 9, but in terms of the smallest* m *satisfying Equation (7) below instead of* (5)*,*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^\ell_S(\langle w, \varphi(\cdot)\rangle) \right] + C^\ell_{\mathsf{dc}} \cdot \sqrt{\frac{d}{m}} \le \varepsilon, \tag{7}$$

where C^ℓ_{dc} *is the loss-specific constant from Equation (6).*

Theorem 11 *For any* $\mathcal{H}$*,* $\varepsilon > 0$ *and Lipschitz or bounded loss* ℓ*,*

$$\mathsf{Lin}^\ell_\varepsilon(\mathcal{H}) \le \mathsf{gLin}^\ell_\varepsilon(\mathcal{H}) \qquad \text{and} \qquad \Omega\left(\frac{\mathsf{dc}^\ell_\varepsilon(\mathcal{H})}{\varepsilon^2}\right) \le \mathsf{gLin}^\ell_\varepsilon(\mathcal{H}) \le \mathcal{O}\left(\frac{\mathsf{dc}^\ell_{\varepsilon/2}(\mathcal{H})}{\varepsilon^2}\right)$$

and analogously for $\mathsf{Lin}_\varepsilon^{\mathcal{D},\ell}$*,* $\mathsf{gLin}_\varepsilon^{\mathcal{D},\ell}$ *and* $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}$ *and any distributions* $\mathcal{D}$ *over* $\mathcal{X}$*.*

The proof of Theorem 11 is presented in Appendix C. Thus, $\mathsf{dc}_\varepsilon(\mathcal{H})$ (and $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H})$) precisely captures "the sample complexity of learning $\mathcal{H}$ using a linear embedding by relying on a guarantee that follows from dimension based generalization bounds", and are therefore *sufficient* for linear learning. In Section 3.3, we will return to the question of whether they are also necessary for the weaker notion of linear learning of Definition 9, i.e. whether they also lower bound Lin_ε and $\mathsf{Lin}_\varepsilon^{\mathcal{D}}$. But before that, we introduce the analogous notions for kernel based learning.

3.2. Kernel Learning Complexity

Recall that for any $\mathscr{D}$, any bounded embedding with $\|\varphi\|_\infty \leq 1$, any R and any Lipschitz loss (c.f. Shalev-Shwartz and Ben-David, 2014),

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in\mathcal{B}(\mathbb{H};R)}\left(\mathcal{L}^\ell_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)-\mathcal{L}^\ell_S(\langle w,\varphi(\cdot)\rangle)\right)\right] \;\leq\; C^\ell_{\mathsf{mc}}\cdot\frac{R^2}{\sqrt{m}}, \tag{8}$$

where C^ℓ_{mc} is twice the Lipschitz constant, which motivates the norm constrained ERM:

$$\textsc{Erm}^\ell_\varphi(S;R) \;:=\; \operatorname*{argmin}_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^\ell_S(\langle w,\varphi(\cdot)\rangle)\,. \tag{9}$$

We therefore define the Kernel Learning Complexity and the Guaranteed Kernel Learning Complexity analogously to Definitions 9 and 10 but relying on $\textsc{Erm}^\ell_\varphi(S;R)$. We must be a bit more careful though, when considering margin based binary classification since neither the 0/1 error nor the margin error are Lipschitz. We can still discuss the ERM w.r.t. the margin loss, but can only use it to bound the population 0/1 loss.

Definition 12 *Fix a hypothesis class $\mathcal{H}\subseteq\mathcal{Y}^\mathcal{X}$, a Lipschitz loss ℓ and parameter $\varepsilon>0$.*

Distributional Kernel Learning Complexity $\mathsf{Ker}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H})$*, parameterized by a distribution $\mathcal{D}$ over $\mathcal{X}$, is the smallest m for which there exists a distribution $\mathcal{P}$ over embeddings $\varphi:\mathcal{X}\to\mathbb{H}$ with $\|\varphi\|_\infty\leq 1$ and a parameter R such that for all realizable distributions $\mathscr{D}$ over $\mathcal{X}\times\mathcal{Y}$ with marginal $\mathcal{D}$ over $\mathcal{X}$,*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in\textsc{Erm}^\ell_\varphi(S;R)}\mathcal{L}^\ell_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\leq\; \varepsilon\,. \tag{10}$$

Kernel Learning Complexity $\mathsf{Ker}^\ell_\varepsilon(\mathcal{H})$ *is the smallest m for which there exists a distribution $\mathcal{P}$ over embeddings $\varphi:\mathcal{X}\to\mathbb{H}$ with $\|\varphi\|_\infty\leq 1$ and a parameter R such that for all realizable distributions $\mathscr{D}$ over $\mathcal{X}\times\mathcal{Y}$, Equation (10) above holds.*

For $\mathcal{Y}=\{1,-1\}$ and $\ell=\ell_{\mathrm{mgn}}$, we define $\mathsf{Ker}_\varepsilon(\mathcal{H}):=\mathsf{Ker}^{\ell_{\mathrm{mgn}}}_\varepsilon(\mathcal{H})$ *and* $\mathsf{Ker}^\mathcal{D}_\varepsilon(\mathcal{H}):=\mathsf{Ker}^{\mathcal{D},\ell_{\mathrm{mgn}}}_\varepsilon(\mathcal{H})$ *analogously, but require that Equation (11) below holds instead of* (10)*:*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in\textsc{Erm}^{\ell_{\mathrm{mgn}}}_\varphi(S;R)}\mathcal{L}^{\ell_{0\text{-}1}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\leq\; \varepsilon\,. \tag{11}$$

As we did in the case of linear learning, to relate $\mathsf{Ker}_\varepsilon(\mathcal{H})$ to $\mathsf{mc}_\varepsilon(\mathcal{H})$, we again consider a stronger notion that requires learning that can be guaranteed based only on the norm, using Equation (8):

Definition 13 *For a Lipschitz loss ℓ, the* **Guaranteed Kernel Learning Complexity** $\mathsf{gKer}^\ell_\varepsilon(\mathcal{H})$ *and* **Distributional Guaranteed Kernel Learning Complexity** $\mathsf{gKer}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H})$ *are defined as in Definition 9, but in terms of the smallest m satisfying Equation (12) below instead of* (10)*,*

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^\ell_S(\langle w,\varphi(\cdot)\rangle)\right]+C^\ell_{\mathsf{mc}}\cdot\frac{B}{\sqrt{m}} \;\leq\; \varepsilon. \tag{12}$$

For $\mathcal{Y} = \{1, -1\}$ and $\ell = \ell_{\mathrm{mgn}}$, $\mathsf{gKer}_\varepsilon(\mathcal{H})$ and $\mathsf{gKer}^{\mathcal{D}}_\varepsilon(\mathcal{H})$ are analogous but we require Equation (13) holds instead:

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}_S^{\ell_{\mathrm{mgn}}}(\langle w, \varphi(\cdot)\rangle) \right] + 2\cdot\frac{B}{\sqrt{m}} \;\leq\; \varepsilon. \tag{13}$$

Theorem 14 *For any $\mathcal{H}$, $\varepsilon > 0$ and Lipschitz or bounded loss ℓ,*

$$\mathsf{Ker}^\ell_\varepsilon(\mathcal{H}) \;\leq\; \mathsf{gKer}^\ell_\varepsilon(\mathcal{H}) \qquad \text{and} \qquad \Omega\left(\frac{\mathsf{mc}^\ell_\varepsilon(\mathcal{H})^2}{\varepsilon^2}\right) \;\leq\; \mathsf{gKer}^\ell_\varepsilon(\mathcal{H}) \;\leq\; \mathcal{O}\left(\frac{\mathsf{mc}^\ell_{\varepsilon/2}(\mathcal{H})^2}{\varepsilon^2}\right)$$

and analogously for $\mathsf{Ker}^{\mathcal{D},\ell}_\varepsilon$, $\mathsf{gKer}^{\mathcal{D},\ell}_\varepsilon$ and $\mathsf{mc}^{\mathcal{D},\ell}_\varepsilon$ for all distributions $\mathcal{D}$ over $\mathcal{X}$.

The proof of Theorem 14 is presented in Appendix C. Thus, $\mathsf{mc}_\varepsilon(\mathcal{H})$ and $\mathsf{mc}^{\mathcal{D}}_\varepsilon(\mathcal{H})$ precisely captures "the sample complexity of learning $\mathcal{H}$ using a kernel with a guarantee that follows from norm based generalization bounds", both for margin-based binary classification, and with respect to a Lipschitz loss.

Remark. Our definitions of Lin_ε and Ker_ε capture realizable learning. We can also consider agnostic variants where we allow any $\mathscr{D}$ and the right hand side of (5), (7), (10), (11), (12) and (13) changes to $\inf_{h\in\mathcal{H}} \mathcal{L}_{\mathscr{D}}(h) + \varepsilon$, for loss functions where this makes sense. The lower bounds on learning of course still hold, and for typical loss functions, including those discussed in this work, we can still get upper bounds in terms of the approximate dimensional and margin complexities.

3.3. Lower Bounds on Learning

We saw that $\mathsf{dc}_\varepsilon(\mathcal{H})$ and $\mathsf{mc}_\varepsilon(\mathcal{H})$ precisely capture $\mathsf{gLin}_\varepsilon(\mathcal{H})$ and $\mathsf{gKer}_\varepsilon(\mathcal{H})$ i.e. "learning based on dimension or norm guarantees". But what about $\mathsf{Lin}_\varepsilon(\mathcal{H})$ and $\mathsf{Ker}_\varepsilon(\mathcal{H})$? Perhaps for specific feature maps, e.g. if the image $\varphi(\mathcal{X})$ is degenerate in special ways, ERM on linear predictors, or perhaps low norm predictors, could give learning guarantees with significantly less than d or R^2 samples? Can we say that $\mathsf{dc}_\varepsilon(\mathcal{H})$ and $\mathsf{mc}_\varepsilon(\mathcal{H})$ also tightly capture $\mathsf{Lin}_\varepsilon(\mathcal{H})$ and $\mathsf{Ker}_\varepsilon(\mathcal{H})$? While we are not able to say this in the distribution-independent setting, we can prove lower bounds in terms of the distribution dependent notion $\mathsf{dc}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H})$.

Theorem 15 *For all $\mathcal{H}$, losses ℓ, distributions $\mathcal{D}$ over $\mathcal{X}$ and $\varepsilon > 0$,*

$$\mathsf{Lin}^\ell_\varepsilon(\mathcal{H}) \;\geq\; \mathsf{Lin}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H}) \;\geq\; \mathsf{dc}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H}) \qquad \textit{and} \qquad \mathsf{Ker}^\ell_\varepsilon(\mathcal{H}) \;\geq\; \mathsf{Ker}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H}) \;\geq\; \mathsf{dc}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H})$$

This follows as a consequence of the *Representer Theorem*, which allows us to replace any high-dimensional embedding by an m dimensional one that is obtained as the span of the embeddings of the samples from $\mathscr{D}$. The proof is presented in Appendix C.

Since Theorem 15 holds for any distribution $\mathcal{D}$, the lower bound on distribution independent learning can also be stated as

$$\mathsf{Lin}^\ell_\varepsilon(\mathcal{H}),\; \mathsf{Ker}^\ell_\varepsilon(\mathcal{H}) \;\geq\; \sup_{\mathcal{D}} \mathsf{dc}^{\mathcal{D}}_\varepsilon(\mathcal{H}). \tag{14}$$

This supremum, which following Theorem 15 tightly characterizes $\sup_{\mathcal{D}} \mathsf{Lin}^{\mathcal{D},\ell}_\varepsilon(\mathcal{H})$, should not be confused with the distribution independent $\mathsf{dc}_\varepsilon(\mathcal{H})$. We can view $\sup_{\mathcal{D}} \mathsf{Lin}^{\mathcal{D}}_\varepsilon(\mathcal{H})$ as corresponding to a semi-supervised learning model where we have unlimited amount of unlabeled data, from which we can infer $\mathcal{D}$, and use it to decide on a distribution over embeddings φ.

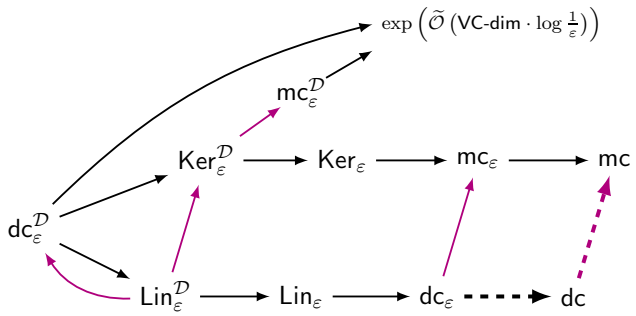

Figure 1: A comparison of all measures introduced, specialized to $\ell_{0\text{-}1}/\ell_{\text{mgn}}$. A solid arrow $A \to B$ denotes $A(\mathcal{H}) \le B(\mathcal{H})$, a solid purple arrow $A \to B$ denotes $A(\mathcal{H}) \le B(\mathcal{H})$ up to some change of parameter ε and some multiplicative factors (either $\text{poly}(1/\varepsilon)$ or $\log(|\mathcal{H}||\mathcal{X}|$ in case of $\mathsf{dc} \to \mathsf{mc}$). $A \dashrightarrow B$ denotes $A(\mathcal{H}) \le B(\mathcal{H})$ and that there exists a class $\mathcal{H}$ for which $A(\mathcal{H}) \ll B(\mathcal{H})$. If A is a distribution-dependent measure and B is a distribution independent measure, then an arrow from $A \to B$ is meant for all $\mathcal{D}$.

Alternate Learning Rules The learning rule we studied as a "kernel method" was to minimize the loss subject to a constraint on the norm, $\min \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle)$ subject to $\|w\|_{\mathbb{H}} \le R$. This is reasonable as it corresponds to our generalization bounds, but often in practice other Pareto-optimal choices are considered, such as the minimum norm zero error (i.e. hard margin) predictor $\min \|w\|_{\mathbb{H}}$ subject to $\mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle) = 0$, or perhaps a more relaxed version, $\min \|w\|_{\mathbb{H}}$ subject to $\mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle) \le \varepsilon$ or Tikhonov-type regularization $\min \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle) + \lambda\|w\|_{\mathbb{H}}$.

All of the above are variants of $\operatorname{argmin}_{w\in\mathbb{H}} g(\mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle), \|w\|_{\mathbb{H}})$ for some monotone function $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}\cup\{\infty\}$, and hence the Representer Theorem holds for all them. Thus $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ would continue to be a lower bound on $\mathsf{Ker}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ for any variant of its definition based on any of the above learning rules.

4. Lower bounds on Probabilistic Distributional Dimension Complexity

In Theorem 15, we established that the sample complexity of learning a hypothesis class $\mathcal{H}$ with dimension-based or kernel-based linear learning is lower bounded by its probabilistic distributional dimension complexity, $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$. In this section, we prove lower bounds on $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ in the case of squared-loss and the zero-one loss, demonstrating the utility of our proposed complexity measures in characterizing the limitations of linear learning.

4.1. Probabilistic dimension complexity w.r.t. Square Loss

Notations. For a distribution $\mathcal{D}$ over $\mathcal{X}$, for any $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ we define $\langle f, g\rangle_{\mathcal{D}} := \mathbb{E}_{x\sim\mathcal{D}} f(x)g(x)$ and $\|f\|_{\mathcal{D}} := \sqrt{\langle f, f\rangle_{\mathcal{D}}} = \sqrt{\mathbb{E}_{x\sim\mathcal{D}} f(x)^2}$. We say that a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is *normalized* if $\|h\|_{\mathcal{D}} = 1$ for every $h \in \mathcal{H}$. For any subset of hypotheses $\mathcal{H}' \subseteq \mathcal{H}$, define its corresponding Gram matrix $G_{\mathcal{H}'}^{\mathcal{D}} \in \mathbb{R}^{|\mathcal{H}'|\times|\mathcal{H}'|}$ as $G_{\mathcal{H}'}^{\mathcal{D}}(h, h') := \langle h, h'\rangle_{\mathcal{D}}$. For any $M \in \mathbb{R}^{t\times p}$ with $t \le p$, we use $\sigma_1(M) \le \ldots \le \sigma_t(M)$ to denote its singular values. For any symmetric $M \in \mathbb{R}^{t\times t}$, we use $\lambda_1(M) \le \ldots \le \lambda_t(M)$ to denote its eigenvalues. We use $\lambda_{\min}(M)$ to mean $\lambda_1(M)$.

Definition 16 (SQ dimension) *For a distribution $\mathcal{D}$ over $\mathcal{X}$, the* Statistical Query dimension *of a normalized hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, denoted as $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H})$, is the largest t for which there exist hypotheses $h_1, \ldots, h_t \in \mathcal{H}$ such that $\langle h_i, h_j \rangle_{\mathcal{D}} \leq 1/2t$ for each $i \neq j$.*

While the Statistical Query dimension is a well studied quantity in learning theory (Blum et al., 1994), we introduce a new measure that is more suited to our goal of proving lower bounds on $\mathsf{dc}_{\varepsilon}^{\mathcal{D},\ell_{\mathrm{sq}}}$. This measure is lower bounded by $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H})$, but in general can be much larger.

Definition 17 (minEV dimension) *For a distribution $\mathcal{D}$ over $\mathcal{X}$, the* min-Eigenvalue dimension *of a normalized hypothesis class $\mathcal{H}$, denoted as $\mathsf{minEV\text{-}dim}^{\mathcal{D}}(\mathcal{H}; \lambda)$, is the largest t for which there exists a subset of hypotheses $H_t := \{h_1, \ldots, h_t\} \in \mathcal{H}$ such that $\lambda_{\min}(G^{\mathcal{D}}_{H_t}) \geq \lambda$.*

Proposition 18 *For all distributions $\mathcal{D}$ over $\mathcal{X}$ and all normalized hypothesis classes $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H}) = t$ implies that $\mathsf{minEV\text{-}dim}^{\mathcal{D}}(\mathcal{H}; 1/2) \geq t$*

Proof Let $\mathcal{H}_t = \{h_1, \ldots, h_t\} \subseteq \mathcal{H}$ such that $\langle h_i, h_j \rangle_{\mathcal{D}} \leq 1/2t$. Thus, all off-diagonal entries of $G^{\mathcal{D}}_{\mathcal{H}_t}$ are at most $1/2t$ in magnitude, whereas all diagonal entries are 1. It follows from Geršgorin (1931) "circle theorem" that all eigenvalues of $G^{\mathcal{D}}_{\mathcal{H}_t}$ are at least $1 - t/2t = 1/2$. ■

Remark. More generally, we could define $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H}; \gamma)$ with respect to parameter $\gamma < 1$, as the largest t for which there exist hypotheses $h_1, \ldots, h_t \in \mathcal{H}$ such that $\langle h_i, h_j \rangle_{\mathcal{D}} \leq \gamma$ for each $i \neq j$. Proposition 18 could then be $\mathsf{SQ\text{-}dim}(\mathcal{H}; \gamma) = t$ implies that $\mathsf{minEV\text{-}dim}(\mathcal{H}; 1 - t\gamma) \geq t$.

Theorem 19 *For all $\varepsilon > 0$, all distributions $\mathcal{D}$ over $\mathcal{X}$ and normalized hypothesis classes $\mathcal{H} \in \mathbb{R}^{\mathcal{X}}$, it holds for any $\lambda \in (2\varepsilon, 1]$ that*

$$\mathsf{dc}_{\varepsilon}^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}) \;\geq\; \left(1 - \frac{2\varepsilon}{\lambda}\right) \cdot \mathsf{minEV\text{-}dim}^{\mathcal{D}}(\mathcal{H}; \lambda)$$

Observe that the bound becomes vacuous at $\varepsilon = \frac{1}{2}$, and rightly so, because the zero function incurs a square loss of $1/2$ for any $h \in \mathcal{H}$, since $\mathcal{H}$ is a normalized hypothesis class. The constant 0 function is realizable with an embedding of dimension 1.

Our proof of Theorem 19 is inspired by the technique due to Alon et al. (2013) for lower bounding the "approximate rank" of a matrix that is well studied in communication complexity. We present the full proof in Appendix D.1. Combining Proposition 18 with Theorem 19 immediately gives us the following corollary.

Corollary 20 *For all distributions $\mathcal{D}$ over $\mathcal{X}$ and normalized hypothesis classes $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$,*

$$\mathsf{dc}_{\varepsilon}^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}) \;\geq\; (1 - 4\varepsilon) \cdot \mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H})\,.$$

4.1.1. APPLICATIONS OF THEOREM 19

We now discuss some applications of our Theorem 19 and Corollary 20.

Example 1 : Parities. Let $\mathcal{X}_n = \{1, -1\}^n$ and $\mathcal{H}^{\oplus}_n = \left\{\chi_S(x) := \prod_{i \in S} x_i : S \subseteq [n]\right\}$ be the class of all parity functions on n bits. Let $\mathcal{D}$ be the uniform distribution over $\mathcal{X}$. For any two distinct subsets $S, T \subseteq [n]$, we have that $\langle \chi_S, \chi_T \rangle_{\mathcal{D}} = 0$. Thus, $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H}^{\oplus}_n) = 2^n$. More strongly, we also have $\mathsf{minEV\text{-}dim}(\mathcal{H}^{\oplus}_n; 1) = 2^n$. Thus, from Theorem 19, we get

$$\mathsf{dc}_{\varepsilon}^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}^{\oplus}_n) \;\geq\; (1 - 2\varepsilon) \cdot 2^n\,.$$

Example 2 : ReLU with bounded weights. The *Rectified Linear Unit* is a popular activation function used in neural networks; given by $x \mapsto [x]_+ = \max\{x, 0\}$. It was recently shown by Yehudai and Shamir (2019) that random features cannot be used to learn (or even approximate) a single ReLU neuron (over standard Gaussian inputs in $\mathbb{R}^n$ with $\mathrm{poly}(n)$ weights) unless the number of features or the magnitude of the learnt coefficients are exponential in n. Using Corollary 20, we are able to improve on this result by removing the restriction on the magnitude of learnt coefficients and obtain a lower bound simply on the number of random features required (this was conjectured to be possible by Yehudai and Shamir (2019)).

Let $\mathcal{H}^{\mathrm{relu}}_{n,W,B} := \{x \mapsto [\langle w, x\rangle + b]_+ : w \in \mathbb{R}^n, b \in \mathbb{R}, \text{ s.t. } \|w\|_2 \le W, |b| \le B\}$ be the class of all functions obtained as a ReLU applied on a linear function with bounded weights.

Theorem 21 (Strengthens Thm 4.2 in Yehudai and Shamir (2019)) *For $\mathcal{D}$ being the standard Gaussian distribution over $\mathbb{R}^n$, there exists a choice of $W \le O(n^3)$ and $B \le O(n^4)$, such that, for any $\varepsilon < 1/4$ that*

$$\mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}^{\mathrm{relu}}_{n,W,B}) \;\ge\; \exp(\Omega(n))$$

Our proof builds on a proposition from Yehudai and Shamir (2019) and also follows the outline there quite closely. However, we believe that this way of presenting the proof is more insightful as it is modular, involving a lower bound on SQ-dimension. The details are deferred to Appendix E.

Example 3 : studied by Allen-Zhu and Li (2019, 2020). Recently, Allen-Zhu and Li (2019, 2020) exhibited functions classes that can provably be "efficiently" learnt using a neural network, but require "large" number of samples or run-time for any kernel method to learn with respect to square loss. In our terminology, the function classes they consider can be shown to have "large" $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{sq}}}$ measure using Theorem 19 and Corollary 20. Since, the function classes they consider are somewhat specialized, we skip the details.

4.2. Probabilistic dimension complexity w.r.t. 0-1 loss

In the previous subsection we considered regression problems, and learning with respect to the squared loss. We now turn to the classification and learning with respect to the 0/1 loss.

We prove a lower bound on the probabilistic distributional dimension complexity w.r.t. $\ell_{0\text{-}1}$ loss for the class of all 1-sparse predictors $\mathcal{H}_n^{1\text{-sp}} \subseteq \{1,-1\}^{\mathcal{X}_n}$ for $\mathcal{X}_n = \{1,-1\}^n$ defined as $\mathcal{H}_n^{1\text{-sp}} := \{h_i : \mathcal{X}_n \to \{1,-1\} : i \in [n] \text{ and } h_i(x) = x_i\}$.

Theorem 22 *Fix $\varepsilon < 1/2$. For $\mathcal{D}$ being the uniform distribution over $\mathcal{X}_n = \{1,-1\}^n$ it holds that,*

$$\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}_n^{1\text{-sp}}) \ge n \cdot \left(\frac{(1-h(\varepsilon))}{4\log(16e/(1-h(\varepsilon)))}\right) - o(n)$$

where $h(q) := q\log_2\left(\frac{1}{q}\right) + (1-q)\log_2\left(\frac{1}{1-q}\right)$ is the binary entropy function.

In particular, we have that $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}_n^{1\text{-sp}}) \ge \Omega(n)$ for any $\epsilon < \frac{1}{2}$, while the bound rightly becomes vacuous at $\varepsilon = \frac{1}{2}$. Contrast this linear scaling with n to the VC dimension of 1-sparse predictors $\mathsf{VC\text{-}dim}(\mathcal{H}_n) \le \log n$, which implies sparse linear predictors are learnable, using a direct approach, which only $O(\log n)$ samples. Thus, Theorem 22 establishes that linear or kernel-based learning would require exponentially more samples than a direct approach.

Theorem 22 also shows that the exponential dependence in our upper bound of $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H})$ in terms of $\mathsf{VC\text{-}dim}(\mathcal{H})$ (Lemma 8) is indeed necessary, and Lemma 8 is, in this sense, tight.

The key technique used in the proof of Theorem 22 is the fact that random $n \times n$ sign-matrices require a sign-rank of $\Omega(n)$ to be even approximated on a constant ($> 1/2$) fraction of the entries. We partition the $n \times 2^n$ sign matrix $M_{\mathcal{H}_n^{\text{1-sp}}}$ randomly into blocks of $n \times n$ matrices and argue that most of those blocks must incur large error if the dimension of the embedding is small. The proof details are deferred to Appendix D.2.

4.2.1. A COMPLEXITY-THEORETIC BARRIER

In Theorem 22 we proved a lower bound on $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}_n)$ for the class of 1-sparse predictors, which has $|\mathcal{X}_n| = 2^{|\mathcal{H}_n|}$. Even just representing a single instance in this example requires $\log|\mathcal{X}_n| = n$ bits, and so the *runtime* for any learning algorithm would also be at least $\Omega(n)$. That is, even though we showed the sample complexity for linear or kernel based learning is exponential in the VC-dimension, i.e. insisting on linear or kernel based learning causes an exponential increase in sample complexity, the sample complexity of linear learning is still no more than linear in the *runtime* or even *memory* of a direct approach. This is in contrast to the examples of Section 4.1, where the lower bound on the sample complexity of linear or kernel based learning was exponential also in the representational cost of instances, i.e. in $\log|\mathcal{X}|$.

Can we prove such a stronger lower bound also with respect to the 0/1 loss, i.e. a lower bound on $\mathsf{dc}_\varepsilon^{\mathcal{D}}$ that is exponential (or even just super-polynomial) in both $\mathsf{VC\text{-}dim}(\mathcal{H})$ and $\log|\mathcal{X}|$? In particular, can we prove a $\mathrm{poly}(n)$ lower bound on $\mathsf{dc}_\varepsilon^{\mathcal{D}}$ for the class of all parities over n bits, for which we do have a strong lower bound w.r.t. square loss?

In turns out that proving such a lower bounds for any explicit class $\mathcal{H}$ will have significant complexity theoretic consequences. Suppose for example, we have an explicit class $\mathcal{H} \subseteq \{1,-1\}^{\mathcal{X}}$ for which we could prove, for some value of $\varepsilon > 0$, that

$$\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}) \;\geq\; (\log|\mathcal{H}||\mathcal{X}|)^{\omega(1)} \cdot \frac{1}{\varepsilon}\,.$$

That is, we could establish a lower bound on $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H})$ that is super-polynomial in $\log|\mathcal{X}|$ and in $\mathsf{VC\text{-}dim}(\mathcal{H})$ (recall that $\mathsf{VC\text{-}dim}(\mathcal{H}) \leq \log|\mathcal{H}|$). As shown by Alman and Williams (2017) (see Lemma 25 & Proposition 24) it will follow that depth-2 threshold circuits computing $E_{\mathcal{H}} : (h,x) \mapsto h(x)$ require size that is at least $(\log|\mathcal{H}||\mathcal{X}|)^{\omega(1)}$, for any binary encoding of $\mathcal{H}$ and $\mathcal{X}$.

Proving super-polynomial lower bounds on the size of depth-2 threshold circuits is a major frontier in Complexity Theory (the best lower bounds known so far is due to Kane and Williams (2016), who show a lower bound of $\widetilde{\Omega}(n^{1.5})$ for an explicit n-bit function). And so, establishing strong lower bounds on linear or kernel based learning with respect to the 0/1 loss for specific classes seems difficult. This explains, perhaps, why recent work on the relative power of deep learning over kernel method focused on regression w.r.t. the square loss, and indicates that establishing similar results also for classification might not be so easy.

Since proving explicit lower bounds for $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H})$ faces a complexity theoretic barrier, we could ask for lower bounds on $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\text{hinge}}}(\mathcal{H})$. Interestingly, it was shown by Balcan et al. (2008) (stated in our notations) that $\mathsf{mc}_\varepsilon^{\mathcal{D},\ell_{\text{hinge}}}(\mathcal{H}) \geq (\frac{2}{\pi} - \varepsilon) \cdot \Omega\left(\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H})^{1/2}\right)$, which suggests the following open question.

Open Question. Can we prove lower bounds on $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\text{hinge}}}(\mathcal{H})$ in terms of $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H})$?

5. Summary

We formalized a notion of Linear Learning ($\mathsf{Lin}^{\ell}_{\varepsilon}$) and Kernel Learning ($\mathsf{Ker}^{\ell}_{\varepsilon}$) with respect to any loss ℓ. We defined probabilistic variants of the classic notions of dimensional complexity ($\mathsf{dc}^{\ell}_{\varepsilon}$) and margin complexity ($\mathsf{mc}^{\ell}_{\varepsilon}$), which we show are equivalent to a notion of "guaranteed" Linear Learning ($\mathsf{gLin}^{\ell}_{\varepsilon}$) and Kernel Learning ($\mathsf{gKer}^{\ell}_{\varepsilon}$) respectively, where the guarantee follows from standard generalization bounds which follow from dimension-based or norm-based arguments respectively. For each of the notions above, we also defined a *distributional version*, where we fix a marginal distribution $\mathcal{D}$ over the input space $\mathcal{X}$.

We showed that $\mathsf{dc}^{\ell}_{\varepsilon}$ and $\mathsf{mc}^{\ell}_{\varepsilon}$ (resp. $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon}$ and $\mathsf{mc}^{\mathcal{D},\ell}_{\varepsilon}$) are *sufficient* for learning with finite dimension or with finite norm embeddings (respectively in the distribution dependent setting). Moreover, in the case of $\ell = \ell_{0\text{-}1}$ loss, $\mathsf{dc}^{\ell_{0\text{-}1}}_{\varepsilon}$ can be exponentially smaller than the classic notion of $\mathsf{dc}^{\ell_{0\text{-}1}}$. We also showed that the distributional versions $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon}$ and $\mathsf{mc}^{\mathcal{D},\ell}_{\varepsilon}$ are upper bounded in terms of the VC-dimension.

Finally, we showed that $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon}$ is *necessary* for learning with either finite dimension or with finite norm embeddings, in the distribution dependent setting and hence also in the distribution independent setting. These connections are summarized in Figure 1.

In the case of $\ell = \ell_{\mathrm{sq}}$, we proved a lower bound $\mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}$ in terms of the notion of $\mathsf{minEV\text{-}dim}^{\mathcal{D}}$, which in turn is lower bounded by $\mathsf{SQ\text{-}dim}^{\mathcal{D}}$; this allows us to re-prove (and even improve upon) similar lower bounds proved in literature (Yehudai and Shamir, 2019; Allen-Zhu and Li, 2019, 2020). In the case of $\ell = \ell_{0\text{-}1}$, we prove a lower bound on $\mathsf{dc}^{\mathcal{D},\ell_{0\text{-}1}}_{\varepsilon}$ of $\Omega(n)$ for the class of 1-sparse predictors on n variables. But this is only logarithmic in $|\mathcal{X}|$. However, we identified a complexity theoretic barrier, namely that any lower bound on $\mathsf{dc}^{\mathcal{D},\ell_{0\text{-}1}}_{\varepsilon}$ for any $\mathcal{D}$ that is super-polynomial in $\log(|\mathcal{H}||\mathcal{X}|)$ for any explicit class $\mathcal{H}$ will imply super-polynomial lower bounds for depth-2 threshold circuits which is long-standing open question in circuit complexity.

We hope that our notions of probabilistic dimensional and margin complexity prove useful in the further understanding of the limitations of linear and kernel learning.

Acknowledgments

We thank Josh Alman, Shai Ben-David, Avrim Blum, Brian Bullins, Surbhi Goel, Mika Göös, Suriya Gunasekar, Adam Klivans, Nati Linial, Raghu Meka, Prasad Raghavendra, Sasha Razborov, Ohad Shamir, Sasha Sherstov, Blake Woodworth and Gilad Yehudai for helpful discussions. We would especially like to thank Surbhi for suggesting the formulation in Corollary 20 in terms of SQ dimension and Mika for suggesting the proof of Theorem 22.

Research was partially supported by NSF BIGDATA award 1546500 and NSF IIS/RI award 1764032. Part of the work was done when the authors were visiting the Simons Institute as part of the program on *Foundations of Deep Learning*.

References

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *arXiv*, abs/1905.10337, 2019. URL http://arxiv.org/abs/1905.10337.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv*, abs/2001.04413, 2020. URL https://arxiv.org/abs/2001.04413.

Josh Alman and R. Ryan Williams. Probabilistic rank and matrix rigidity. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 641–652, 2017. URL https://doi.org/10.1145/3055399.3055484.

Noga Alon, Troy Lee, Adi Shraibman, and Santosh S. Vempala. The approximate rank of a matrix and its algorithmic applications: approximate rank. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 675–684, 2013. URL https://doi.org/10.1145/2488608.2488694.

Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 47–80. JMLR.org, 2016. URL http://proceedings.mlr.press/v49/alon16.html.

Rosa I. Arriaga and Santosh S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 616–623, 1999. doi: 10.1109/SFFCS.1999.814637. URL https://doi.org/10.1109/SFFCS.1999.814637.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008. URL https://doi.org/10.1007/s10994-008-5059-5.

Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half spaces. *Journal of Machine Learning Research*, 3(Nov):441–461, 2002. URL http://jmlr.org/papers/v3/bendavid02a.html.

Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using fourier analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 253–262, 1994. URL https://doi.org/10.1145/195058.195147.

Arkadev Chattopadhyay and Nikhil S. Mande. A Short List of Equalities Induces Large Sign Rank. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 47–58, 2018. URL https://doi.org/10.1109/FOCS.2018.00014.

Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2422–2430, 2017. URL http://papers.nips.cc/paper/6836-sgd-learns-the-conjugate-kernel-class-of-the-network.

Jürgen Forster and Hans Ulrich Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theoretical Computer Science*, 350(1):40–48, 2006. URL https://doi.org/10.1016/j.tcs.2005.10.015.

Jürgen Forster, Niels Schmitt, Hans Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003. URL https://doi.org/10.1023/A:1022905618164.

Semyon Aronovich Geršgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, pages 749–754, 1931. URL http://mi.mathnet.ru/izv5235.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of combinatorial theory. Series A*, 69(2):217–232, 1995.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 8580–8589, 2018. URL http://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.

G. J. O. Jameson. *Summing and Nuclear Norms in Banach Space Theory*. London Mathematical Society Student Texts. Cambridge University Press, 1987. URL https://doi.org/10.1017/CBO9780511569166.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Daniel M. Kane and Ryan Williams. Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 633–643, 2016. doi: 10.1145/2897518.2897636. URL https://doi.org/10.1145/2897518.2897636.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184, 2007. URL http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1313–1320, 2008. URL http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomizati

Alexander A Razborov and Alexander A Sherstov. The sign-rank of AC^0. *SIAM Journal of Computing*, 39(5):1833–1855, 2010. URL https://doi.org/10.1137/080744037.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132. URL https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html.

Alexander A. Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008. URL https://doi.org/10.1007/s00037-008-0242-4.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005*, pages 545–560, 2005. URL https://doi.org/10.1007/11503415_37.

Nathan Srebro, Noga Alon, and Tommi S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1321–1328, 2004. URL http://papers.nips.cc/paper/2700-generalization-error-bounds-for-collaborative-prediction-with-low-rank-matrices.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 6594–6604, 2019. URL http://papers.nips.cc/paper/8886-on-the-power-and-limitations-of-random-features-for-understanding-neural-networ

Appendix A. Relating dc and mc : Proof of Lemma 5

Proof of Lemma 5 For any Hilbert space $\mathbb{H}$, by the lemma of (Johnson and Lindenstrauss, 1984), we have that there exists a distribution $\mathcal{A}$ over projections $\pi : \mathbb{H} \to \mathbb{R}^d$ such that for any $u, v \in \mathbb{H}$,

$$\Pr_{\pi\sim\mathcal{A}}\left[\left|\langle u,v\rangle_{\mathbb{H}} - \langle \pi(u),\pi(v)\rangle_{\mathbb{R}^k}\right| > \tau\right] < \delta \qquad \text{for } d = \Theta\left(\frac{\|u\|_{\mathbb{H}}^2\|v\|_{\mathbb{H}}^2}{\tau^2}\log\frac{1}{\delta}\right). \tag{15}$$

We can also derive an expectation version of the above to get

$$\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left|\langle u,v\rangle_{\mathbb{H}} - \langle \pi(u),\pi(v)\rangle_{\mathbb{R}^k}\right|^2 \le \mathcal{O}\left(\frac{\|u\|_{\mathbb{H}}^2\|v\|_{\mathbb{H}}^2}{d}\right) \tag{16}$$

which also implies

$$\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left|\langle u,v\rangle_{\mathbb{H}} - \langle \pi(u),\pi(v)\rangle_{\mathbb{R}^k}\right| \le \mathcal{O}\left(\frac{\|u\|_{\mathbb{H}}\|v\|_{\mathbb{H}}}{\sqrt{d}}\right) \tag{17}$$

Let $\mathcal{P}_{\mathrm{mc}}$ be a distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{H}$ with $\|\varphi\|_\infty \le 1$ that realizes the definition of $\mathsf{mc}_\varepsilon^\ell(\mathcal{H}) =: R$. That is, for all distributions $\mathcal{D}$ over $\mathcal{X}$ and all $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}_{\mathrm{mc}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^\ell_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)\right] \le \varepsilon\,.$$

Consider a distribution $\mathcal{P}_{\mathrm{dc}}$ over embeddings $\psi : \mathcal{X} \to \mathbb{R}^d$ obtained as $\psi(x) = \pi(\varphi(x))$ for independently sampled $\varphi \sim \mathcal{P}_{\mathrm{mc}}$ and $\pi \sim \mathcal{A}$. For any distribution $\mathcal{D}$ over $\mathcal{X}$ and any $h \in \mathcal{H}$, we have,

$$\begin{aligned}\mathop{\mathbb{E}}_{\psi\sim\mathcal{P}_{\mathrm{dc}}}\left[\inf_{w\in\mathbb{R}^d}\mathcal{L}^\ell_{\mathcal{D},h}(\langle w,\psi(\cdot)\rangle)\right] &\le \mathop{\mathbb{E}}_{\substack{\varphi\sim\mathcal{P}_{\mathrm{mc}}\\ \pi\sim\mathcal{A}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^\ell_{\mathcal{D},h}(\langle \pi(w),\pi(\varphi(\cdot))\rangle)\right] \\ &\le \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}_{\mathrm{mc}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\mathcal{L}^\ell_{\mathcal{D},h}(\langle \pi(w),\pi(\varphi(\cdot))\rangle)\right] \end{aligned} \tag{18}$$

Proof of (i). We first infer from (15) that for any $u, v \in \mathbb{H}$,

$$\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left[\mathbb{1}\left\{\langle\pi(u),\pi(v)\rangle < 0\right\}\right] \;\leq\; \mathbb{1}\left\{\langle u,v\rangle < \tau\right\} + \delta \qquad \text{for } d = \Theta\left(\frac{\|u\|_{\mathbb{H}}^2\|v\|_{\mathbb{H}}^2}{\tau^2}\log\frac{1}{\delta}\right) \tag{19}$$

Starting from the inner term in (18), for any $w \in \mathbb{H}$ with $\|w\|_{\mathbb{H}} \leq R$ and $\|\varphi\|_\infty \leq 1$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\mathcal{L}^{\ell_{0\text{-}1}}_{\mathcal{D},h}(\langle\pi(w),\pi(\varphi(\cdot))\rangle) \;&=\; \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\mathbb{1}\left\{\langle\pi(w),\pi(\varphi(x))\rangle\, h(x) \leq 0\right\} \\
&\leq\; \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathbb{1}\left\{\langle w,\varphi(x)\rangle\, h(x) \leq 1\right\} + \eta \qquad \ldots\text{(from (19))} \\
&=\; \mathcal{L}^{\ell_{\mathrm{mgn}}}_{\mathcal{D},h}(\langle\pi(w),\pi(\varphi(\cdot))\rangle) + \eta
\end{aligned}$$

where we instantiate (19) with $\tau = 1$, $\delta = \eta$, by setting $d = O(R^2\log(1/\eta))$. Plugging this upper bound into (18), we get our desired goal

$$\mathop{\mathbb{E}}_{\psi\sim\mathcal{P}_{\mathrm{dc}}}\left[\inf_{w\in\mathbb{R}^d}\mathcal{L}^{\ell_{0\text{-}1}}_{\mathcal{D},h}(\langle w,\psi(\cdot)\rangle)\right] \;\leq\; \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}_{\mathrm{mc}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^{\ell_{\mathrm{mgn}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)\right] + \eta \;\leq\; \varepsilon + \eta\,.$$

Proof of (ii). We use (17). For any $w \in \mathbb{H}$ with $\|w\|_{\mathbb{H}} \leq R$, we have from L-Lipschitzness of ℓ and $\|\varphi\|_\infty \leq 1$ that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left[\mathcal{L}^{\ell}_{\mathcal{D},h}(\langle\pi(w),\pi(\varphi(\cdot))\rangle)\right] - \mathcal{L}^{\ell}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle) \\
&=\; \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left[\ell(\langle\pi(w),\pi(\varphi(x))\rangle, h(x)) - \ell(\langle w,\varphi(x)\rangle, h(x))\right] \\
&\leq\; L\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left|\langle\pi(w),\pi(\varphi(x))\rangle - \langle w,\varphi(x)\rangle\right| \\
&\leq\; \mathcal{O}\left(\frac{LR}{\sqrt{d}}\right)
\end{aligned}$$

Combining this with (18), we get,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\psi\sim\mathcal{P}_{\mathrm{dc}}}\left[\inf_{w\in\mathbb{R}^d}\mathcal{L}^{\ell}_{\mathcal{D},h}(\langle w,\psi(\cdot)\rangle)\right] \;&\leq\; \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}_{\mathrm{mc}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^{\ell}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle) + \mathcal{O}\left(\frac{LR}{\sqrt{d}}\right)\right] \\
&\leq\; \varepsilon + \mathcal{O}\left(\frac{LR}{\sqrt{d}}\right)
\end{aligned}$$

Thus, we get our desired statement for a choice of $d = \mathcal{O}\left(LR/\eta\right)^2$.

Proof of (iii). We use (16) and (17). We use (17). For any $w \in \mathbb{H}$ with $\|w\|_{\mathbb{H}} \leq R$ we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left[\mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle\pi(w),\pi(\varphi(\cdot))\rangle)\right] - \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)\\
&= \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left[(h(x)-\langle\pi(w),\pi(\varphi(x))\rangle)^2 - (h(x)-\langle w,\varphi(x)\rangle)^2\right]\\
&\leq \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left|h(x)-\langle w,\varphi(x)\rangle\right|\cdot\left|\langle\pi(w),\pi(\varphi(x))\rangle - \langle w,\varphi(x)\rangle\right|\\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\mathop{\mathbb{E}}_{\pi\sim\mathcal{A}}\left|\langle\pi(w),\pi(\varphi(x))\rangle - \langle w,\varphi(x)\rangle\right|^2\\
&\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left|h(x)-\langle w,\varphi(x)\rangle\right|\cdot\mathcal{O}\left(\frac{R}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{R^2}{d}\right)\\
&\leq \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left|h(x)-\langle w,\varphi(x)\rangle\right|^2\right)^{1/2}\cdot\mathcal{O}\left(\frac{R}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{R^2}{d}\right)\\
&= \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)^{1/2}\cdot\mathcal{O}\left(\frac{R}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{R^2}{d}\right)
\end{aligned}$$

Combining this with (18), we get,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\psi\sim\mathcal{P}_{\mathrm{dc}}}\left[\inf_{w\in\mathbb{R}^d}\mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\psi(\cdot)\rangle)\right]\\
&\leq \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}_{\mathrm{mc}}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)}\mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle) + \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)^{1/2}\cdot\mathcal{O}\left(\frac{R}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{R^2}{d}\right)\right]\\
&\leq \varepsilon + \mathcal{O}\left(\frac{\sqrt{\varepsilon}R}{\sqrt{d}}\right) + \mathcal{O}\left(\frac{R^2}{d}\right)
\end{aligned}$$

where, we use that $\mathbb{E}_\varphi \inf_w \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)^{1/2} \leq \left(\mathbb{E}_\varphi \inf_w \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)\right)^{1/2} \leq \sqrt{\varepsilon}$. Thus, we get our desired statement for a choice of $d = R^2\cdot\mathcal{O}\left((\varepsilon+\eta)/\eta^2\right)$. This completes the proof for all the parts (i), (ii) and (iii). The analogous cases relating $\mathsf{dc}^{\mathcal{D},\ell}_{\varepsilon+\eta}$ and $\mathsf{mc}^{\mathcal{D},\ell}_{\varepsilon}$ follows similarly. ■

Appendix B. Proofs of Separation between Deterministic and Probabilistic Dimension Complexity

B.1. Exponential gap : Proof of Theorem 6

We first introduce a variant of probabilistic dimension complexity that requires a stronger point-wise notion of correctness.

Definition 23 *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and a loss ℓ and a parameter $\varepsilon \geq 0$. The* point-wise probabilistic dimension complexity $\mathsf{dc}^{\mathrm{pt},\ell}_{\varepsilon}(\mathcal{H})$ *is the smallest d for which there exists a distribution $\mathcal{P}$ over a pair of embeddings $(\varphi : \mathcal{X}\to\mathbb{R}^d, w : \mathcal{H}\to\mathbb{R}^d)$ such that,*

$$\sup_{(x,h)\in\mathcal{X}\times\mathcal{H}}\mathop{\mathbb{E}}_{(\varphi,w)\sim\mathcal{P}}\left[\ell(\langle w(h),\varphi(x)\rangle, h(x))\right] \leq \varepsilon\,.$$

This notion of point-wise probabilistic dimension complexity requires that (the distribution over) w is chosen without the knowledge of the distribution $\mathcal{D}$ over $\mathcal{X}$ and hence is stronger than probabilistic dimension complexity as in Definition 2. In particular, we have the following.

Proposition 24 *For all $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, loss ℓ and parameter $\varepsilon > 0$, it holds that,*

$$\sup_{\mathcal{D}} \mathsf{dc}_{\varepsilon}^{\mathcal{D},\ell}(\mathcal{H}) \;\leq\; \mathsf{dc}_{\varepsilon}^{\ell}(\mathcal{H}) \;\leq\; \mathsf{dc}_{\varepsilon}^{\mathrm{pt},\ell}(\mathcal{H})$$

The notion of $\mathsf{dc}_{\varepsilon}^{\mathrm{pt},\ell_{0\text{-}1}}(\mathcal{H})$ is equivalent to the notion of *probabilistic sign-rank* studied in the communication complexity. In particular, stating in our notations, Alman and Williams (2017) showed that if the function $E_{\mathcal{H}} : \mathcal{H} \times \mathcal{X} \to \{1,-1\}$ given by $E_{\mathcal{H}}(h,x) := h(x)$ is computable by small depth-2 threshold circuits (for any encoding of $\mathcal{H}$ and $\mathcal{X}$ into bits), then $\mathsf{dc}_{\varepsilon}^{\mathrm{pt},\ell_{0\text{-}1}}(\mathcal{H})$ is also small.

Lemma 25 (Alman and Williams (2017)) *If $E_{\mathcal{H}}$ is computable by a depth-2 threshold circuit of size s, then*

$$\mathsf{dc}_{\varepsilon}^{\mathrm{pt},\ell_{0\text{-}1}}(\mathcal{H}) \leq O\left(\frac{s^2 \log^2(|\mathcal{H}| \cdot |\mathcal{X}|)}{\varepsilon}\right)$$

Theorem 6 now follows readily from a recent lower bound on sign-rank shown by Chattopadhyay and Mande (2018) for matrices that are computable by small depth-2 threshold circuits.

Proof of Theorem 6 We describe the construction of the class $\mathcal{H}$, which is indexed by $\{1,-1\}^n$. To describe how an $h \in \mathcal{H}$ *acts on* an $x \in \mathcal{X}$, we divide the n bits in h and x into k blocks by writing $h = (h_1, \ldots, h_k)$ and $x = (x_1, \ldots, x_k)$ where each $h_i, x_i \in \{1,-1\}^p$ with $kp = n$. The hypothesis h on input x outputs -1 iff the largest index $i \in [k]$ for which $h_i = x_i$ holds is an odd index.[1] For $p = k^{1/3} + \log k$, it was shown by Chattopadhyay and Mande (2018) that

$$\mathsf{dc}(\mathcal{H}) \geq 2^{\Omega(n^{1/4})}.$$

Chattopadhyay and Mande (2018) also observe that $E_{\mathcal{H}} : (h,x) \mapsto h(x)$ is computable by a depth-2 threshold circuit of size $O(n)$. Thus, from Lemma 25, we have that

$$\mathsf{dc}_{\varepsilon}^{\mathrm{pt},\ell_{0\text{-}1}}(\mathcal{H}) \;\leq\; O\left(\frac{n^4}{\varepsilon}\right)$$

Combining with Proposition 24 we get our desired separation. ■

B.2. "Infinite" gap : Proof of Theorem 7

We first prove Lemma 8 that probabilistic distributional dimension complexity can be upper bounded in terms of VC dimension.

Proof of Lemma 8 A classic result due to Haussler (1995) shows that for any distribution $\mathcal{D}$ over $\mathcal{X}$ there exists a cover $\mathcal{C}_\varepsilon \subseteq \mathcal{H}$, with $|\mathcal{C}_\varepsilon| \leq c \cdot \mathsf{VC\text{-}dim}(\mathcal{H}) \cdot (K/\varepsilon)^{\mathsf{VC\text{-}dim}(\mathcal{H})}$ for some universal constants c, K, such that,

$$\forall h \in \mathcal{H},\ \exists c_h \in \mathcal{C}_\varepsilon \text{ such that } \Pr_{x \sim \mathcal{D}}[h(x) \neq c_h(x)] \leq \varepsilon\,.$$

1. In communication complexity parlance, the associated $M_{\mathcal{H}}$ would be called a "pattern matrix".

Thus for any given distribution $\mathcal{D}$, we can construct a (deterministic) embedding $\varphi : \mathcal{X} \to \mathbb{R}^{|\mathcal{C}_\varepsilon|}$ given as $\varphi(x) = (c(x))_{c\in\mathcal{C}_\varepsilon}$ and $w(h) = (\mathbb{1}[c = c_h])_{c\in\mathcal{C}_\varepsilon}$ satisfying the property that,

$$\forall h \in \mathcal{H} \;:\; \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathbb{1}[h(x) \neq \operatorname{sign}(\langle\varphi(x), w(h)\rangle)] \leq \varepsilon.$$

This implies that $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}) \leq |\mathcal{C}_\varepsilon|$. Note that, since $\langle w(h), \varphi(x)\rangle$ always takes values in $\{1, -1\}$, $\mathsf{dc}_{2\varepsilon}^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H})$ and $\mathsf{dc}_{2\varepsilon}^{\mathcal{D},\ell_{\mathrm{hinge}}}(\mathcal{H})$ are also at most $|\mathcal{C}_\varepsilon|$.

Also, observe that if we can scale φ by $1/\sqrt{|\mathcal{C}_\varepsilon|}$, we will have $\|\varphi\|_\infty \leq 1$. To compensate for this, we can scale up w by $\sqrt{|\mathcal{C}_\varepsilon|}$ and get the desired upper bound on $\mathsf{mc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$. ■

Proof of Theorem 7 Alon et al. (2016) showed that for $\mathcal{X} = \{1, -1\}^n$ there exists a hypothesis class $\mathcal{H} \subseteq \{1, -1\}^{\mathcal{X}}$ such that $\mathsf{VC\text{-}dim}(\mathcal{H}) = 2$ but $\mathsf{dc}(\mathcal{H}) \geq 2^{\Omega(n)}$. Note that $\mathsf{dc}^{\ell_{\mathrm{sq}}}(\mathcal{H})$ and $\mathsf{dc}^{\ell_{\mathrm{hinge}}}(\mathcal{H})$ are each larger than $\mathsf{dc}(\mathcal{H})$. Also note that $\mathsf{mc}(\mathcal{H}) \geq \Omega(\sqrt{\mathsf{dc}(\mathcal{H})/n})$ (from the classic result relating mc and dc). Thus we get the desired lower bound on $\mathsf{mc}(\mathcal{H})$, $\mathsf{mc}^{\ell_{\mathrm{sq}}}(\mathcal{H})$ and $\mathsf{mc}^{\ell_{\mathrm{hinge}}}(\mathcal{H})$ as well. On the other hand, from Lemma 8, we get that both $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ (for $\ell \in \{\ell_{0\text{-}1}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}$) and $\mathsf{mc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ (for $\ell \in \{\ell_{\mathrm{mgn}}, \ell_{\mathrm{sq}}, \ell_{\mathrm{hinge}}\}$) are at most $\mathcal{O}\left(1/\varepsilon^2\right)$ for every distribution $\mathcal{D}$ over $\mathcal{X}$. ■

Appendix C. Proofs of Upper and Lower Bounds on Learning

C.1. Learning via Random embeddings : Proof of Theorems 11 and 14

Proof of Theorem 11 $\mathsf{Lin}_\varepsilon^\ell(\mathcal{H}) \leq \mathsf{gLin}_\varepsilon^\ell(\mathcal{H})$ and $\Omega\left(\frac{\mathsf{dc}_\varepsilon^\ell(\mathcal{H})}{\varepsilon^2}\right) \leq \mathsf{gLin}_\varepsilon^\ell(\mathcal{H}) \leq \mathcal{O}\left(\frac{\mathsf{dc}_{\varepsilon/2}^\ell(\mathcal{H})}{\varepsilon^2}\right)$

Let $\mathcal{P}$ be the distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{R}^d$ underlying the definition of $\mathsf{gLin}_\varepsilon^\ell(\mathcal{H}) =: m$. That is, we have for any realizable distribution $\mathscr{D}$ over $\mathcal{X} \times \mathcal{Y}$ that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\inf_{w\in\mathbb{R}^d} \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle)\right] + C_{\mathsf{dc}}^\ell \cdot \sqrt{\frac{d}{m}} \;\leq\; \varepsilon\,. \tag{20}$$

On the other hand, from standard generalization bounds (cf. Equation (6)), we have for any choice of $\varphi : \mathcal{X} \to \mathbb{R}^d$ and $\mathscr{D}$ that

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\sup_{w\in\mathbb{R}^d} \mathcal{L}_{\mathscr{D}}^\ell(\langle w, \varphi(\cdot)\rangle) - \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle)\right] \;\leq\; C_{\mathsf{dc}}^\ell \cdot \sqrt{\frac{d}{m}}\,.$$

And hence,

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\sup_{w\in\mathrm{ERM}_\varphi^\ell(S)} \mathcal{L}_{\mathscr{D}}^\ell(\langle w, \varphi(\cdot)\rangle)\right] \;\leq\; \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\inf_{w\in\mathbb{R}^d} \mathcal{L}_S^\ell(\langle w, \varphi(\cdot)\rangle)\right] + C_{\mathsf{dc}}^\ell \sqrt{\frac{d}{m}}$$

Thus, taking expectation over $\varphi \sim \mathcal{P}$, we have from (20) that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m} \left[\sup_{w\in\mathrm{ERM}_\varphi^\ell(S)} \mathcal{L}_{\mathscr{D}}^\ell(\langle w, \varphi(\cdot)\rangle)\right] \;\leq\; \varepsilon$$

Thus, we get $\mathsf{Lin}^{\ell}_{\varepsilon}(\mathcal{H}) \le m = \mathsf{gLin}^{\ell}_{\varepsilon}(\mathcal{H})$. It also follows that $\mathsf{dc}_{\varepsilon}(\mathcal{H}) \le \varepsilon^2 \mathsf{gLin}^{\ell}_{\varepsilon}(\mathcal{H})$, since firstly $d \le \varepsilon^2 m$ by definition of $\mathsf{gLin}^{\ell}_{\varepsilon}(\mathcal{H}) = m$. Moreover, if we let $\mathscr{D}$ to be the distribution sampled as $x \sim \mathcal{D}$ and $y = h(x)$ for some $h \in \mathcal{H}$, we get,

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell}_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle)\right] \;\le\; \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in \textsc{Erm}^{\ell}_{\varphi}(S)} \mathcal{L}^{\ell}_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle)\right] \;\le\; \varepsilon$$

Finally, it remains to show that $\mathsf{gLin}^{\ell}_{\varepsilon}(\mathcal{H}) \le O(\mathsf{dc}^{\ell}_{\varepsilon/2}(\mathcal{H})/\varepsilon^2)$. Let $\mathcal{P}$ be the distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{R}^d$ that realizes the definition of $\mathsf{dc}^{\ell}_{\varepsilon/2}(\mathcal{H}) =: d$. Thus, we have for any realizable distribution $\mathscr{D}$ over $\mathcal{X} \times \mathcal{Y}$ that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell}_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle)\right] \;\le\; \frac{\varepsilon}{2}\,. \tag{21}$$

Now, for any choice of $\varphi : \mathcal{X} \to \mathbb{R}^d$ and any $w_* \in \mathbb{R}^d$ we have

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell}_{S}(\langle w, \varphi(\cdot)\rangle)\right] \;\le\; \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\mathcal{L}^{\ell}_{S}(\langle w_*, \varphi(\cdot)\rangle)\right] \;=\; \mathcal{L}^{\ell}_{\mathscr{D}}(\langle w_*, \varphi(\cdot)\rangle)$$

Taking infimum over w_* (in RHS) and an expectation over $\varphi \sim \mathcal{P}$, we get,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell}_{S}(\langle w, \varphi(\cdot)\rangle)\right] + C^{\ell}_{\mathsf{dc}} \cdot \sqrt{\frac{d}{m}} \\
&\le \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell}_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle)\right] + C^{\ell}_{\mathsf{dc}} \cdot \sqrt{\frac{d}{m}} \\
&\le \frac{\varepsilon}{2} + C^{\ell}_{\mathsf{dc}} \cdot \sqrt{\frac{d}{m}} \qquad \ldots \text{(from (21))} \\
&\le \varepsilon \qquad \ldots \text{(for a choice of } m = \mathcal{O}(d/\varepsilon^2))
\end{aligned}$$

This establishes $\mathsf{gLin}^{\ell}_{\varepsilon}(\mathcal{H}) \le \mathcal{O}(\mathsf{dc}^{\ell}_{\varepsilon/2}(\mathcal{H})/\varepsilon^2)$, thereby completing the proof for the distribution-independent case. The distribution-dependent analogs follow in an identical manner. ■

Proof of Theorem 14 $\mathsf{Ker}^{\ell}_{\varepsilon}(\mathcal{H}) \le \mathsf{gKer}^{\ell}_{\varepsilon}(\mathcal{H})$ and $\Omega\left(\frac{\mathsf{mc}^{\ell}_{\varepsilon}(\mathcal{H})^2}{\varepsilon^2}\right) \le \mathsf{gKer}^{\ell}_{\varepsilon}(\mathcal{H}) \le \mathcal{O}\left(\frac{\mathsf{mc}^{\ell}_{\varepsilon/2}(\mathcal{H})^2}{\varepsilon^2}\right)$

This proof is very similar to that of Theorem 11, except that we use norm-based generalization bounds instead of dimension-based ones. We present the proof for $\ell_{0\text{-}1}/\ell_{\mathrm{mgn}}$ and the case of general Lipshitz ℓ follows in a similar manner.

Let $\mathcal{P}$ be the distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{H}$ underlying the definition of $\mathsf{gKer}_{\varepsilon}(\mathcal{H}) =: m$. That is, we have for any realizable distribution $\mathscr{D}$ over $\mathcal{X} \times \mathcal{Y}$ that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\mathrm{mgn}}}_{S}(\langle w, \varphi(\cdot)\rangle)\right] + C_{\mathsf{mc}} \cdot \frac{R}{\sqrt{m}} \;\le\; \varepsilon\,. \tag{22}$$

On the other hand, from standard norm based generalization bounds (see Equation (8)), we have for any choice of $\varphi : \mathcal{X} \to \mathbb{H}$ and $\mathscr{D}$ that

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathscr{D}}(\langle w, \varphi(\cdot)\rangle) - \mathcal{L}^{\ell_{\mathrm{mgn}}}_{S}(\langle w, \varphi(\cdot)\rangle)\right] \;\le\; C_{\mathsf{mc}} \cdot \frac{R}{\sqrt{m}}\,.$$

And hence,

$$\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in \textsc{Erm}^{\ell_{\text{mgn}}}_{\varphi}(S;R)} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\le\; \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\text{mgn}}}_{S}(\langle w,\varphi(\cdot)\rangle)\right] + C_{\mathsf{mc}}\frac{R}{\sqrt{m}}$$

Thus, taking expectation over $\varphi\sim\mathcal{P}$, we have from (22) that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in \textsc{Erm}^{\ell_{0\text{-}1}}_{\varphi}(S)} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\le\; \varepsilon$$

Thus, we get $\mathsf{Ker}_\varepsilon(\mathcal{H}) \le m = \mathsf{gKer}_\varepsilon(\mathcal{H})$. It also follows that $\mathsf{mc}_\varepsilon(\mathcal{H}) \le \varepsilon\sqrt{\mathsf{gLin}_\varepsilon(\mathcal{H})}$, since firstly $R \le \varepsilon\sqrt{m}$ by definition of $\mathsf{gKer}_\varepsilon(\mathcal{H}) = m$. Moreover, if we let $\mathscr{D}$ to be the distribution sampled as $x\sim\mathcal{D}$ and $y = h(x)$ for some $h\in\mathcal{H}$, we get,

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathcal{D},h}(\langle w,\varphi(\cdot)\rangle)\right] \;\le\; \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\sup_{w\in \textsc{Erm}^{\ell_{0\text{-}1}}_{\varphi}(S;R)} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\le\; \varepsilon$$

Finally, it remains to show that $\mathsf{gKer}_\varepsilon(\mathcal{H}) \le O(\mathsf{mc}_{\varepsilon/2}(\mathcal{H})/\varepsilon^2)$. Let $\mathcal{P}$ be the distribution over embeddings $\varphi:\mathcal{X}\to\mathbb{H}$ with $\|\varphi\|_\infty \le 1$ that realizes the definition of $\mathsf{mc}_{\varepsilon/2}(\mathcal{H}) =: R$. Thus, we have for any realizable distribution $\mathscr{D}$ over $\mathcal{X}$ that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\text{mgn}}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] \;\le\; \frac{\varepsilon}{2}\,. \tag{23}$$

Now, for any choice of $\varphi:\mathcal{X}\to\mathbb{H}$ and any $w_*\in\mathbb{H}$ with $\|w_*\|_{\mathbb{H}}\le R$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\text{mgn}}}_{S}(\langle w,\varphi(\cdot)\rangle)\right] &\le \mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\mathcal{L}^{\ell_{\text{mgn}}}_{S}(\langle w_*,\varphi(\cdot)\rangle)\right]\\ &= \mathcal{L}^{\ell_{\text{mgn}}}_{\mathscr{D}}(\langle w_*,\varphi(\cdot)\rangle)\end{aligned}$$

Finally, taking expectation over $\varphi\sim\mathcal{P}$ and taking infimum over w_* (in RHS), we get

$$\begin{aligned}&\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\mathop{\mathbb{E}}_{S\sim\mathscr{D}^m}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\text{mgn}}}_{S}(\langle w,\varphi(\cdot)\rangle)\right] + C_{\mathsf{mc}}\cdot\frac{R}{\sqrt{m}}\\ &\le \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}}\left[\inf_{w\in\mathcal{B}(\mathbb{H};R)} \mathcal{L}^{\ell_{\text{mgn}}}_{\mathscr{D}}(\langle w,\varphi(\cdot)\rangle)\right] + C_{\mathsf{mc}}\cdot\frac{R}{\sqrt{m}}\\ &\le \frac{\varepsilon}{2} + C_{\mathsf{mc}}\cdot\frac{R}{\sqrt{m}} \quad \ldots\text{(from (23))}\\ &\le \varepsilon \qquad \ldots\text{(for a choice of } m = \mathcal{O}(R^2/\varepsilon^2))\end{aligned}$$

This establishes $\mathsf{gKer}_\varepsilon(\mathcal{H}) \le \mathcal{O}(\mathsf{mc}_{\varepsilon/2}(\mathcal{H})/\varepsilon^2)$, thereby completing the proof for the distribution-independent case. The distribution-dependent analogs follow in an identical manner. ∎

C.2. Lower Bound on Learning : Proof of Theorem 15

Proof of Theorem 15 We start with part (i). The first inequality of $\mathsf{Lin}_\varepsilon^\ell(\mathcal{H}) \geq \mathsf{Lin}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H})$ holds by definition; we focus on the second inequality. Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X}$ and $\varepsilon > 0$. Let $\mathcal{P}$ be the distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{R}^d$ that realizes the definition of $\mathsf{Lin}_\varepsilon(\mathcal{H}) =: m$ for some d. For any $h \in \mathcal{H}$, let $\mathscr{D}_h$ be the distribution over $\mathcal{X} \times \mathcal{Y}$ given by $(x, h(x))$ for $x \sim \mathcal{D}$ (that is, $\mathscr{D}_h$ is a distribution *realizable* under $\mathcal{H}$). Thus, we have for any $h \in \mathcal{H}$ that

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathscr{D}_h^m} \left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^\ell_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle) \right] \leq \varepsilon.$$

For any $\varphi : \mathcal{X} \to \mathbb{R}^d$ and $S \sim \mathscr{D}_h^m$, define the subspace spanned by embedding of the data $U_{\varphi,S} := \operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_m)\}$. We show that $\textsc{Erm}^\ell_\varphi(S) \cap U_{\varphi,S} \neq \emptyset$; also known as "Representer Theorem". Namely, for any $w \in \textsc{Erm}^\ell_\varphi(S)$, we can decompose $w = w^{||} + w^\perp$ such that $w^{||} \in U_{\varphi,S}$ and $\langle w^\perp, u\rangle = 0$ for all $u \in U_{\varphi,S}$. Thus, $\langle w, \varphi(x)\rangle = \langle w^{||}, \varphi(x)\rangle$ for each $x \in S$. Hence $w^{||} \in \textsc{Erm}^\ell_\varphi(S) \cap U_{\varphi,S}$. Thus, we have

$$\mathop{\mathbb{E}}_{\substack{\varphi\sim\mathcal{P}\\ S\sim\mathscr{D}_h^m}} \left[\inf_{w\in U_{\varphi,S}} \mathcal{L}^\ell_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle) \right] \leq \mathop{\mathbb{E}}_{\substack{\varphi\sim\mathcal{P}\\ S\sim\mathscr{D}_h^m}} \left[\inf_{w\in \textsc{Erm}^\ell_\varphi(S)\cap U_{\varphi,S}} \mathcal{L}^\ell_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle) \right] \leq \varepsilon.$$

Note that in the definition of $U_{\varphi,S}$, the labels sampled from $\mathscr{D}_h$ are unused. So we abuse notations and define $U_{\varphi,S}$ even for $S \sim \mathcal{D}^m$. In order to show that $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell}(\mathcal{H}) \leq m$ we construct a distribution $\mathcal{P}_{\mathrm{dc}}$ over embeddings $\psi : \mathcal{X} \to \mathbb{R}^m$ as follows: Sample $\varphi \sim \mathcal{P}$ and $S \sim \mathcal{D}^m$ and let $\psi(x) := \pi_{\varphi,S}(\varphi(x))$, where $\pi_{\varphi,S} : \mathbb{R}^d \to \mathbb{R}^m$ is the projection onto the subspace $U_{\varphi,S}$, expressed in terms of some canonical orthonormal basis. Note that for any φ, S and $w \in U_{\varphi,S}$, it holds that $\langle w, \varphi(x)\rangle = \langle \pi_{\varphi,S}(w), \psi(x)\rangle$. Thus, we get

$$\mathop{\mathbb{E}}_{\psi\sim\mathcal{P}_{\mathrm{dc}}} \left[\inf_{w\in\mathbb{R}^m} \mathcal{L}^\ell_{\mathcal{D},h}\left(\langle w, \psi(\cdot)\rangle\right) \right] = \mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \left[\inf_{w\in U_{\varphi,S}} \mathcal{L}^\ell_{\mathcal{D},h}\left(\langle w, \varphi(\cdot)\rangle\right) \right] \leq \varepsilon.$$

Part (ii) follows in an identical manner, so we skip the details. ■

Appendix D. Proofs of Lower Bounds on Probabilistic Distributional Dimension Complexity

D.1. Case of square loss : Proof of Theorem 19

Our proof is inspired by the technique for lower bounding the approximate rank of a matrix due to Alon et al. (2013).

Proof of Theorem 19 For $\lambda > 2\varepsilon$, let $t := \mathsf{minEV\text{-}dim}^{\mathcal{D}}(\mathcal{H}; \lambda)$. That is, we have hypotheses $\mathcal{H}_t = \{h_1, \ldots, h_t\}$ with $\lambda_{\min}(G^{\mathcal{D}}_{\mathcal{H}_t}) \geq \lambda$. Let $d := \mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H})$, that is, there exists a distribution $\mathcal{P}$ over pairs of embeddings $(\varphi : \mathcal{X} \to \mathbb{R}^d, w : \mathcal{H} \to \mathbb{R}^d)$[2] such that for all $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{(\varphi,w)\sim\mathcal{P}} \left[\mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w(h), \varphi(\cdot)\rangle) \right] \leq \varepsilon\,.$$

2. by choosing $w : h \mapsto \arg\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle)$

In particular, if we average over $h \in \mathcal{H}_t$,

$$\mathop{\mathbb{E}}_{(\varphi,w)\sim\mathcal{P}} \mathop{\mathbb{E}}_{\substack{h\sim\mathcal{H}_t \\ x\sim\mathcal{D}}} \ell_{\mathrm{sq}}(\langle w(h), \varphi(x)\rangle, h(x)) \leq \varepsilon\,.$$

Thus, we can fix a deterministic pair of embeddings $(\varphi_* : \mathcal{X} \to \mathbb{R}^d, w_* : \mathcal{H} \to \mathbb{R}^d)$ in the support of $\mathcal{P}$ for which,

$$\mathop{\mathbb{E}}_{\substack{h\sim\mathcal{H}_t \\ x\sim\mathcal{D}}} \ell_{\mathrm{sq}}(\langle w_*(h), \varphi_*(x)\rangle, h(x)) \leq \varepsilon\,. \tag{24}$$

We have $G := G^{\mathcal{D}}_{\mathcal{H}_t} = MM^\top$ where $M \in \mathbb{R}^{t\times\mathcal{X}}$ is given by $M(h,x) := \sqrt{\mathcal{D}(x)} \cdot h(x)$ for all $h \in \mathcal{H}_t$ and $x \in \mathcal{X}$. Since $\lambda_{\min}(G) \geq \lambda$ we have for all $v \in \mathbb{R}^t$ that $v^\top G v \geq \lambda\|v\|_2^2$. In particular, we have

$$\forall v \in \mathbb{R}^t \quad : \quad \|M^\top v\|_2 \geq \sqrt{\lambda}\|v\|_2\,. \tag{25}$$

On the other hand, the embedding pair (φ_*, w_*) defines a rank-d matrix $A \in \mathbb{R}^{t\times\mathcal{X}}$ given by $A(h,x) := \sqrt{\mathcal{D}(x)}\,\langle w_*(h), \varphi_*(x)\rangle$ for each $h \in \mathcal{H}_t$ and $x \in \mathcal{X}$.

We define $E \in \mathbb{R}^{t\times\mathcal{X}}$ as $E(h,x) := M(h,x) - A(h,x)$. We have from (24)

$$\|E\|_F^2 \;=\; \sum_{h\in\mathcal{H}_t} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(\langle w_*(h), \varphi_*(x)\rangle - h(x)\right)^2 \;\leq\; 2\varepsilon t$$

In particular, we get

$$\sum_{i=1}^{t} \sigma_i(E)^2 \leq 2\varepsilon t\,. \tag{26}$$

On the other hand, since $\mathsf{rank}(A) \leq d$, there exists a subspace $S \subseteq \mathbb{R}^t$ of dimension $t-d$, such that $\|A^\top v\|_2 = 0$ for all $v \in S$. By triangle inequality, we get $0 = \|A^\top v\|_2 \geq \|M^\top v\|_2 - \|E^\top v\|_2$. From (25) we have $\|M^\top v\|_2 \geq \sqrt{\lambda}$. Thus, $\|E^\top v\|_2 \geq \sqrt{\lambda}$ for all $v \in S$. From the Courant-Fischer-Weyl min-max theorem, we get $\sigma_t(E) \geq \ldots \geq \sigma_{d+1}(E) \geq \sqrt{\lambda}$. Combining this with (26) implies $(t-d)\lambda \leq 2\varepsilon t$. Finally this implies $\mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}) \geq \mathsf{dc}_\varepsilon^{\mathcal{D},\ell_{\mathrm{sq}}}(\mathcal{H}_t) \geq \left(1 - \frac{2\varepsilon}{\lambda}\right) t$ as desired. ■

D.2. Case of 0-1 loss : Proof of Theorem 22

In order to prove Theorem 22, we use a key fact from Srebro et al. (2004) that provides an upper bound on the number of sign-matrices with sign-rank below a given bound. Namely, let $\mathsf{SM}(n,d)$ be the number of sign-matrices $M \in \{1,-1\}^{n\times n}$ with $\mathsf{sign\text{-}rank}(M) \leq d$.

Lemma 26 (Srebro et al. (2004)) *For all $n \geq k \geq 1$, it holds that* $\mathsf{SM}(n,d) \leq \left(\frac{8en}{d}\right)^{2dn}$.

Proof of Theorem 22 Let $\mathcal{P}$ be the distribution over pair of embeddings $(\varphi : \mathcal{X}_n \to \mathbb{R}^d, w : \mathcal{H}_n^{\text{1-sp}} \to \mathbb{R}^d)$[3] that realizes the definition of $\mathsf{dc}_\varepsilon^{\mathcal{D}}(\mathcal{H}_n^{\text{1-sp}}) =: d$. If we sample h uniformly in $\mathcal{H}_n^{\text{1-sp}}$, we have

$$\mathop{\mathsf{Pr}}_{\substack{x\sim\mathcal{D} \\ h\sim\mathcal{H}_n^{\text{1-sp}}}} [\mathrm{sign}(\langle w(h), \varphi(x)\rangle) \neq h(x)] \leq \varepsilon\,. \tag{27}$$

3. by choosing $w : h \mapsto \arg\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell_{0\text{-}1}}_{\mathcal{D},h}(\langle w, \varphi(\cdot)\rangle)$

On the other hand, consider a random subset $S \subseteq \mathcal{X}_n$ of size $|S| = n$ and the hypothesis class $\mathcal{H}_n^{\text{1-sp}}$ evaluated only on inputs $x \in S$. A key step in this proof is to show that for $\gamma < 1/2$ and $c := d/n$,

$$\Pr_S\left[\Pr_{\substack{x\sim S\\ h\sim\mathcal{H}_n^{\text{1-sp}}}}[\operatorname{sign}(\langle w(h),\varphi(x)\rangle) \neq h(x)] \leq \gamma\right] \;\leq\; 2^{-n^2\left(1-h(\gamma)-2c\log\left(\frac{8e}{c}\right)-o(1)\right)}\,. \tag{28}$$

This follows by a simple counting argument. For any $n \times n$ sign-matrix M and $\gamma < 1/2$, the number of sign-matrices A such that $\Pr_{(i,j)\sim[n]\times[n]}[M(i,j) \neq A(i,j)] \leq \gamma$ is at most $\sum_{r=0}^{\gamma n^2}\binom{n^2}{r} \leq 2^{(h(\gamma)+o(1))n^2}$. From Lemma 26, we have that $\mathsf{SM}(n,d) \leq \left(\frac{8en}{d}\right)^{2dn} = 2^{2c\log\left(\frac{8e}{c}\right)n^2}$ where $c := d/n$. Thus, the number of $n \times n$ sign-matrices that agree with some sign-matrix of sign-rank $\leq d$ on at least $(1-\gamma)$ fraction of the entries is at most $2^{\left(h(\gamma)+2c\log\left(\frac{8e}{c}\right)+o(1)\right)n^2}$.

On the other hand, the number of distinct $n \times n$ sign-matrices obtainable by sampling S is at least $(2^n - n)^n \geq 2^{(1-o(1))n^2}$. Thus, (28) follows.

By linearity of expectation, if we partition $\mathcal{X}_n$ into subsets $S_1, \ldots, S_{2^n/n}$ each of size n, then in expectation, the fraction of S_i's for which

$$\Pr_{\substack{x\sim S_i\\ h\sim\mathcal{H}_n^{\text{1-sp}}}}[\operatorname{sign}(\langle w(h),\varphi(x)\rangle) \neq h(x)] > \gamma$$

holds is at least $1 - 2^{-n^2\left(1-h(\gamma)-2c\log\left(\frac{8e}{c}\right)-o(1)\right)}$. In particular, we can fix such a partition for which this happens. And for such a partition, we get that,

$$\begin{aligned}\Pr_{\substack{x\sim\mathcal{D}\\ h\sim\mathcal{H}_n^{\text{1-sp}}}}[\operatorname{sign}(\langle w(h),\varphi(x)\rangle) \neq h(x)] \;&=\; \Pr_i\Pr_{\substack{x\sim S_i\\ h\sim\mathcal{H}_n^{\text{1-sp}}}}[\operatorname{sign}(\langle w(h),\varphi(x)\rangle) \neq h(x)]\\ &>\; \gamma\cdot\left(1 - 2^{-n^2\left(1-h(\gamma)-2c\log\left(\frac{8e}{c}\right)-o(1)\right)}\right).\end{aligned}$$

Combining this with (27), we get for any choice of γ that

$$\gamma\cdot\left(1 - 2^{-n^2\left(1-h(\gamma)-2c\log\left(\frac{8e}{c}\right)-o(1)\right)}\right) \leq \varepsilon$$

In particular, if we choose $\gamma = \varepsilon/(1-2^{-n})$, we get

$$2^{-n^2\left(1-h(\gamma)-2c\log\left(\frac{8e}{c}\right)-o(1)\right)} \;\geq\; 2^{-n}\,.$$

And hence,

$$2c\log\left(\frac{8e}{c}\right) \;\geq\; 1 - h\left(\frac{\varepsilon}{1-2^{-n}}\right) - \frac{1}{n} - o_n(1) \;\geq\; 1 - h(\varepsilon) - o_n(1)\,.$$

Thus,

$$c \;\geq\; \frac{1-h(\varepsilon)}{4\log(16e/(1-h(\varepsilon)))} - o_n(1)\,.$$

This concludes the proof. ∎

Appendix E. Lower Bounds for ReLU Functions : Proof of Theorem 21

Our proof proceeds in a modular fashion: Instead of directly lower bounding minEV-dim for $\mathcal{H}^{\mathrm{relu}}_{n,W,B}$, we prove a lower bound for the class obtained as linear combination of a $\mathrm{poly}(n)$ number of functions in $\mathcal{H}^{\mathrm{relu}}_{n,W,B}$. Towards this goal, for any class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, define

$$\kappa \cdot \mathcal{H} := \{\kappa h : h \in \mathcal{H}\} \qquad \text{and} \qquad \mathcal{H}^{k,A} := \left\{ \sum_{i=1}^{k} a_i h_i : \sum_i a_i^2 \le A \text{ and } h_i \in \mathcal{H} \right\}$$

Proposition 27 *For all $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, all distribution $\mathcal{D}$ over $\mathcal{X}$, and parameters κ, k, A,*

(i) $\mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon/\kappa}(\mathcal{H}) \;=\; \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\sqrt{\kappa}\cdot\mathcal{H})$ *for all* $t \in \mathbb{R}$

(ii) $\mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\mathcal{H}^{k,A}) \;\le\; \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon/kA}(\mathcal{H})$ *for all* $k \in \mathbb{N}$ *and* $A \in \mathbb{R}$

Thus, combining the two parts,

$$\mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\mathcal{H}^{k,A}) \;\le\; \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\sqrt{kA}\cdot\mathcal{H}) \tag{29}$$

Proof Part (i) follows easily by observing that square loss is quadratic in the scaling of $\mathcal{H}$ (and $w \in \mathbb{R}^d$). To establish Part (ii): Let $\mathcal{P}$ be the distribution over embeddings $\varphi : \mathcal{X} \to \mathbb{R}^d$ that realizes the definition of $\mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon/kA}(\mathcal{H}) =: d$. For any φ and any $g = \sum_{i=1}^{k} a_i h_i \in \mathcal{H}^{k,A}$, we have,

$$\begin{aligned}
&\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},g}(\langle w, \varphi(\cdot)\rangle) \;=\; \frac{1}{2} \inf_{w\in\mathbb{R}^d} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(\sum_{i=1}^{k} a_i h_i(x) - \langle w, \varphi(x)\rangle \right)^2 \\
&= \frac{1}{2} \inf_{w_1,\ldots,w_k\in\mathbb{R}^d} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left(\sum_{i=1}^{k} a_i h_i(x) - \left\langle \sum_i a_i w_i, \varphi(x) \right\rangle \right)^2 \qquad \ldots (\text{setting } w = \sum_i a_i w_i) \\
&\le \frac{k}{2} \cdot \sum_{i=1}^{k} a_i^2 \cdot \inf_{w_i\in\mathbb{R}^d} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} (h_i(x) - \langle w_i, \varphi(x)\rangle)^2 \\
&= k \cdot \sum_{i=1}^{k} a_i^2 \cdot \inf_{w_i\in\mathbb{R}^d} \mathcal{L}^{\ell_{\mathrm{sq}}}_{\mathcal{D},h_i}(\langle w_i, \varphi(\cdot)\rangle)
\end{aligned}$$

The proof concludes by taking an expectation over $\varphi \sim \mathcal{P}$,

$$\mathop{\mathbb{E}}_{\varphi\sim\mathcal{P}} \left[\inf_{w\in\mathbb{R}^d} \mathcal{L}^{\mathcal{D},\ell_{\mathrm{sq}}}_{g}(\langle w, \varphi(\cdot)\rangle) \right] \;\le\; k \sum_i a_i^2 \cdot \frac{\varepsilon}{kA} \;\le\; \varepsilon\,.$$

■

Proof of Theorem 21 We will show a lower bound on the SQ-dimension of a class of linear combinations of ReLU neurons. In order to do, we consider for any odd a, the univariate function

$$\psi_a(z) := -1 + [z+a]_+ + \sum_{i=1}^{a-1} 2\cdot(-1)^i \cdot [z+a-2i]_+ - [z-a]+$$

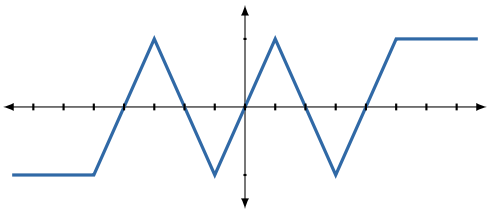

Figure 2: Plot of $\psi_5 : \mathbb{R} \to \mathbb{R}$

See Figure 2 for an illustration of this function. We now consider the class

$$\mathcal{H}_n^{\text{zig}} := \left\{ \psi_a(\langle w, x \rangle) : w \in \mathbb{R}^n, \|w\|_2 = n \text{ for } a = 6n^2 + 1 \right\} .$$

The key idea for showing a lower bound on $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H}_n^{\text{zig}})$ is the following proposition that can be inferred[4] from Proposition 4.2 in Yehudai and Shamir (2019); we skip the details.

Proposition 28 (Prop 4.2 in Yehudai and Shamir (2019)) *There exist constants $c, c' > 0$ such that, for $a = 6n^2 + 1$ and $\mathcal{D}$ being the standard n-variate Gaussian distribution,*

(i) For all $w \in \mathbb{R}^n$ with $\|w\| = n$, it holds that $\|\psi_a(\langle w, x \rangle)\|_{\mathcal{D}} \geq c'$.

(ii) For u, v sampled uniformly at random from $\{w : \|w\| = n\}$,

$$\mathop{\mathbb{E}}_{u,v} \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \psi_a(\langle u, x \rangle) \psi_a(\langle v, x \rangle) \right)^2 \leq \exp(-cn) .$$

Thus, if we sample $u_1, \ldots, u_t$ randomly from $\{w : \|w\| = n\}$, then (via Markov's inequality and a union bound) we will have with probability at least $1/2$ that,

$$\text{for all } i \neq j \quad : \quad \left| \mathbb{E}_{x \sim \mathcal{D}} \ \psi_a(\langle u_i, x \rangle) \psi_a(\langle u_j, x \rangle) \right| \leq t^2 \cdot \exp(-cn) .$$

In particular, for $t := \exp(cn/3)/2$ there exist $u_1, \ldots, u_t \in \mathbb{R}^n$ such that $\|u_i\| = n$ and all pairwise correlations $|\langle \psi_a(\langle u_i, x \rangle), \psi_a(\langle u_j, x \rangle) \rangle_{\mathcal{D}}| \leq \exp(-cn/3)/4 \leq 1/2t$. Thus, we get that, $\mathsf{SQ\text{-}dim}^{\mathcal{D}}(\mathcal{H}_n^{\text{zig}}) \geq \exp(\Omega(n))$. Note however that there is a slight technicality here in that $\mathcal{H}_n^{\text{zig}}$ is not a normalized hypothesis class. But observe that all hypotheses in $\mathcal{H}_n^{\text{zig}}$ have the same norm $\|\cdot\|_{\mathcal{D}}$ which is at least c'. Thus, we can make $\mathcal{H}_n^{\text{zig}}$ normalized by scaling it by $\|\psi_a(\langle u, \cdot \rangle)\|_{\mathcal{D}}^{-1} \leq 1/c'$. This would increase the correlations by a factor of at most $(1/c')^2$. Thus, from Corollary 20, we have that $\mathsf{dc}_{\varepsilon}^{\mathcal{D}, \ell_{\text{sq}}}(\mathcal{H}_n^{\text{zig}}) \geq (1 - 4\varepsilon) \exp(\Omega(n))$.

Observe that every $g \in \mathcal{H}_n^{\text{zig}}$ can be written as a linear combination of $6n^2 + 3$ ReLU neurons of the form $[\langle w, x \rangle + b]_+$, where $\|w\| \leq n$ and $|b| \leq 6n^2 + 1 < 7n^2$ (where we can simulate the constant term with $w = 0$), where each coefficient in the linear combination is at most 2. Thus, in

4. Part (i) is verbatim. For Part (ii), we can first infer the desired claim for a fixed u and a random v, and then take an expectation over u.

our notation, $\mathcal{H}^{\mathrm{zig}}_n \subseteq (\mathcal{H}^{\mathrm{relu}}_{n,n,7n^2})^{k,A}$ for $k = 6n^2+3$ and $A = 4(6n^2+3)$. Thus, we get,

$$\begin{aligned}
\exp(\Omega(n)) &\leq \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\mathcal{H}^{\mathrm{zig}}_n) \\
&\leq \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}((\mathcal{H}^{\mathrm{relu}}_{n,n,7n^2})^{k,A}) \\
&\leq \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\sqrt{kA}\cdot\mathcal{H}^{\mathrm{relu}}_{n,n,7n^2}) \qquad \ldots \text{(from Proposition 27)} \\
&\leq \mathsf{dc}^{\mathcal{D},\ell_{\mathrm{sq}}}_{\varepsilon}(\mathcal{H}^{\mathrm{relu}}_{n,14n^3,98n^4})
\end{aligned}$$

where the last step uses that $\kappa \cdot \mathcal{H}^{\mathrm{relu}}_{n,W,B} = \mathcal{H}^{\mathrm{relu}}_{n,\kappa W,\kappa B}$ (which follows from the homogeneity of ReLU). This completes the proof. ■