Trajectory growth lower bounds for random sparse deep ReLU networks[*]

Ilan Price [‡†] and Jared Tanner [†‡]

Abstract. This paper considers the growth in the length of one-dimensional trajectories as they are passed through deep ReLU neural networks, which, among other things, is one measure of the expressivity of deep networks. We generalise existing results, providing an alternative, simpler method for lower bounding expected trajectory growth through random networks, for a more general class of weights distributions, including sparsely connected networks. We illustrate this approach by deriving bounds for sparse-Gaussian, sparse-uniform, and sparse-discrete-valued random nets. We prove that trajectory growth can remain exponential in depth with these new distributions, including their sparse variants, with the sparsity parameter appearing in the base of the exponent.

Key words. deep learning, random curves, random sparse matrices, expected arc-length, neural network expressivity

AMS subject classifications. 62M45, 60D05, 65F50, 15B52

1. Introduction.

Deep neural networks continue to set new benchmarks for machine learning accuracy across a wide range of tasks, and are the basis for many algorithms we use routinely and on a daily basis. One fundamental set of theoretical questions concerning deep networks relates to their *expressivity*. There remain different approaches to understanding and quantifying neural network expressivity. Some results take a classical approximation theory approach, focusing on the relationship between the architecture of the network and the classes of functions it can accurately approximate ([19, 4, 14, 24]). Another more recent approach has been to apply persistent homology to characterise expressivity ([10]), while [22] focus on global curvature, and the ability of deep networks to disentangle manifolds. Other works concentrate specifically on networks with piecewise linear activation functions, using the number of linear regions ([21]) or the volume of the boundaries between linear regions ([12]) in input space. More generally, geometric notions of expressivity of both trained and random nets has been investigated from multiple perspectives in recent years ([5, 8, 13]). In 2017, [23] proposed trajectory length as a measure of expressivity; in particular, they consider the expected change in length of a one-dimensional trajectory as it is passed through Gaussian random neural networks (see Figure 1 for an illustration). Their primary theoretical result was that, in expectation, the length of a one-dimensional trajectory which is passed through a fully-connected, Gaussian network is *lower bounded* by a factor that is exponential with depth, but not with width.

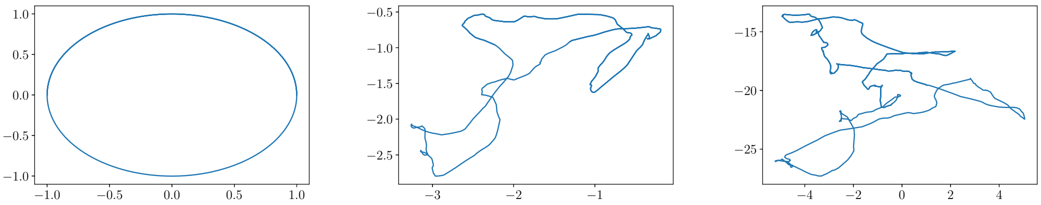
One-dimensional trajectories and their evolution through deep networks are also of interest in their own right because they constitute simple data manifolds. Firstly, we commonly assume that the real data which we aim to correctly classify or predict with a deep network lie on one

[*]Submitted to the editors 21 January 2020.

Funding: This publication is based on work supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

[†]Mathematical Institute, Oxford University (ilan.price@maths.ox.ac.uk, tanner@maths.ox.ac.uk).

[‡]Alan Turing Institute



(a) Input (b) Layer 6 (c) Layer 12

Figure 1: A circular trajectory, passed through a ReLU network with $\sigma_w = 2$. The plots show the pre-activation trajectory at different layers projected down onto 2 dimensions.

or more manifolds, and thus design a network to perform appropriately on such a manifold. Secondly, researchers are beginning to consider whether the *output* (manifolds) of generator networks could be a good model for real word data manifolds, for example, as priors for a variety of inverse problems ([20, 15]). Both of these hypotheses motivate an understanding of how manifolds are acted upon by deep networks.

Our results in this paper pertain specifically to the 'trajectory length' measure of expressivity. We produce a simpler proof than in the pioneering work of [23], which also generalises their results, deriving similar lower bounds for a broader class of random deep neural networks.

Theoretical work of this nature is important because it allows for more straightforward transfer and adaptation of prior theoretical results to new contexts of interest. For example, there is a current surge in research around low-memory networks, training sparse networks, and network pruning. Sparsely connected networks have shown the capacity to retain very high test accuracy ([7, 11]), increased robustness ([2, 1]), with much smaller memory footprints, and less power consumption ([26]). The approach we take in this work enables us to extend results from dense random networks to sparse ones. It also allows us to consider the other weight distributions of sparse-Gaussian, sparse-uniform and sparse-discrete networks (see Definitions 1.2 - 1.4).

More specifically we make **the following contributions**:

1. We provide an alternative, simpler method for lower bounding expected trajectory growth through random networks, for a more general class of weights distributions (Theorem 2.5).
2. We illustrate this approach by deriving bounds for sparse-Gaussian, sparse-uniform, and sparse-discrete random nets. We prove that trajectory growth can be exponential in depth with these distributions, with the sparsity appearing in the base of the exponential (Corollaries 2.2 - 2.4).
3. We observe that the expected length growth factor is strikingly similar across the aforementioned three distributions. This suggests a universality of the expected growth in length for iid centered distributions determined only by the variance and sparsity (Figure 3).

1.1. Notation. We consider feedforward ReLU deep neural networks. We denote a the d-th post-activation layer as $z^{(d)}$, and the subsequent pre-activation layer as $h^{(d)}$, such that

$$h^{(d)} = W^{(d)} z^{(d)} + b^{(d)}, \qquad z^{(d+1)} = \phi(h^{(d)}),$$

where $\phi(x) := \max(x, 0)$ is applied elementwise. We denote $x = z^{(0)}$.

We use $f_{NN}(x; \mathcal{P}, \mathcal{Q})$ to denote a random feedforward deep neural network which takes as input the vector x, and is parameterised by random weight matrices $W^{(d)}$ with entries sampled iid from the distribution $\mathcal{P}$, and bias vectors $b^{(d)}$ with entries drawn iid from distribution $\mathcal{Q}$.

Definition 1.1. *A* ***random sparse network*** *with sparsity parameter α, denoted $f_{NN}(x; \alpha, \mathcal{P}, \mathcal{Q})$, is a random feedforward network in which all weights are sampled from a mixture distribution of the form*

$$w_{ij} \sim \alpha \mathcal{P} + (1 - \alpha)\delta,$$

where δ is the delta distribution at 0, and $\mathcal{P}$ is some other distribution. In other words, weights are 0 with probability $1-\alpha$, and sampled from $\mathcal{P}$ with probability α. Biases are drawn iid from $\mathcal{Q}$.

Definition 1.2. *A* ***sparse-Gaussian network*** *is a random sparse network $f_{NN}(x; \alpha, \mathcal{P}, \mathcal{Q})$, where $\mathcal{P} = \mathcal{N}(0, \sigma_w^2)$ and $\mathcal{Q} = \mathcal{N}(0, \sigma_b^2)$.*

Definition 1.3. *A* ***sparse-uniform network*** *is a random sparse network $f_{NN}(x; \alpha, \mathcal{P}, \mathcal{Q})$, where $\mathcal{P} = \mathcal{U}(-C_w, C_w)$ and $\mathcal{Q} = \mathcal{U}(-C_b, C_b)$.*

Definition 1.4. *A* ***sparse-discrete network*** *is a random sparse network $f_{NN}(x; \alpha, \mathcal{P}, \mathcal{Q})$, where $\mathcal{P}$ is a uniform distribution over a finite, discrete, symmetric set $\mathcal{W}$, with cardinality $|\mathcal{W}| = N_w$, and $\mathcal{Q}$ is a uniform distribution over a finite, discrete, symmetric set $\mathcal{B}$, with cardinality $|\mathcal{B}| = N_b$.*

For a weight matrix W in a random sparse network, with w_i denoting the i^{th} row, we define $w_{\mathcal{P}_i}$ as the vector containing only the $\mathcal{P}$-distributed entries of w_i.

We define a trajectory $x(t)$ in input space as a curve between two points, say x_0 and x_1, parameterized by a scalar $t \in [0, 1]$, with $x(0) = x_0$ and $x(1) = x_1$, and we define $z^{(d)}(x(t)) = z^{(d)}(t)$ to be the image of the trajectory in layer d of the network. The trajectory length $l(x(t))$ is given by the standard arc length,

$$l(x(t)) = \int_t \left|\left| \frac{dx(t)}{dt} \right|\right| dt.$$

As in the work by [23], this paper considers trajectories with $x(t + dt)$ having a non-trivial component perpendicular to $x(t)$ for all t, dt.

Finally, we say a probability density or mass function $f_X(x)$ is even if $f_X(-x) = f_X(x)$ for all random vectors x in the sample space.

2. Expected Trajectory Growth Through Random Networks. [23] considered ReLU and hard-tanh Gaussian networks with the standard deviation scaled by $1/\sqrt{k}$. Their result with respect to ReLU networks is captured in the following theorem.

Theorem 2.1 ([23]). *Let* $f_{NN}(x;\mathcal{N}(0,\sigma_w^2/k),\mathcal{N}(0,\sigma_b^2))$ *be a random Gaussian deep ReLU neural network with layers of width* k*, then*

$$\mathbb{E}[l(z^{(d)}(t))] \geq \mathcal{O}\left(\frac{\sigma_w\sqrt{k}}{\sqrt{k+1}}\right)^d \cdot l(x(t)),$$

for $x(t)$ *a 1-dimensional trajectory in input space.*

There are, however, other network weight distributions which may be of interest. For example, the expressivity and generative power of *sparse* networks are of particular interest in the current moment ([3]), given the current interest in low-memory and low-energy networks, training sparse networks, and network pruning ([7, 11, 26]). We prove that even for sparse random networks, trajectory growth can remain exponential in depth given sufficiently large initialisation scale σ_w. Scaling σ_w by $1/\sqrt{k}$ can yield a width-independent lower bound on this growth. Moreover, a sufficiently high sparsity fraction $(1-\alpha)$ results in a lower bound which, instead of growing exponentially, shrinks exponentially to zero. This is captured by the following result.

Corollary 2.2 (**Trajectory growth in deep sparse-Gaussian random networks**). *Let* $f_{NN}(x;\alpha,\mathcal{N}(0,\sigma_w^2),\mathcal{N}(0,\sigma_b^2))$ *be a sparse-Gaussian, feedforward ReLU network as defined in Section 1.1, with layers of width* k*. Then*

$$\mathbb{E}[l(z^{(d)}(t))] \geq \left(\frac{\alpha\sigma_w\sqrt{k}}{\sqrt{2\pi}}\right)^d \cdot l(x(t)), \tag{2.1}$$

for $x(t)$ *a 1-dimensional trajectory in input space.*

Corollary 2.2 with $\alpha = 1$ and σ_w replaced by $\sigma_w/\sqrt{k}$ recovers a bound which is very similar to the prior bound by [23] in Theorem 2.1.

Beyond Gaussian weights, we consider other distributions commonly used for initialising and analysing deep networks. Uniform distributions, for example, still constitute the default initialisations of linear network layers in both Pytorch and Tensorflow (uniform according to $\mathcal{U}(-1/\sqrt{k},1/\sqrt{k})$ in the case of Pytorch, and uniform according to $\mathcal{U}(-6/\sqrt{k_{in}+k_{out}},6/\sqrt{k_{in}+k_{out}})$ – a.k.a the Glorot/Xavier uniform initialization ([9]) – in the case of Tensorflow). We prove an analogous lower bound for uniformly distributed weights.

Corollary 2.3 (**Trajectory growth in deep sparse-uniform random networks**). *Let* $f_{NN}(x;\alpha,\mathcal{U}(-C_w,C_w),\mathcal{U}(-C_b,C_b))$ *be a sparse-Uniform, feedforward ReLU network as defined in Section 1.1, with layers of width* k*. Then*

$$\mathbb{E}[l(z^{(d)}(t))] \geq \left(\frac{\alpha C_w\sqrt{k}}{4\sqrt{2}}\right)^d \cdot l(x(t)), \tag{2.2}$$

for $x(t)$ *a 1-dimensional trajectory in input space.*

Another research direction which has gathered some momentum in recent years are quantized or discrete-valued deep neural networks ([18, 16, 17]), including recent work using integer valued weights ([25]). This motivates consideration of discrete weight distributions, in addition to continuous ones. As an example of such, we prove a similar lower bound for networks with weights and biases uniformly sampled from finite, symmetric, discrete sets.

Corollary 2.4 (**Trajectory growth in deep sparse-discrete random networks**). *Let $f_{NN}(x;\alpha,\mathcal{P},\mathcal{Q})$ be a sparse-discrete random feedforward ReLU network as defined in Section 1.1, and layers of width k. Then*

$$\mathbb{E}[l(z^{(d)}(t))] \geq \left(\frac{\alpha\sqrt{k}}{2\sqrt{2}}\cdot\frac{\sum_{w\in\mathcal{W}}|w|}{N_w}\right)^d \cdot l(x(t)), \tag{2.3}$$

for $x(t)$ a 1-dimensional trajectory in input space.

In all cases these lower bounds show how to choose the combination of σ_w and α to guarantee (or not) exponential growth in trajectory length in expectation at initialisation.

The main idea behind the derivation of these results is to consider how the length of a small piece of a trajectory (some $\|dz^{(d)}\|$) grows from one layer to the next ($\|dz^{(d+1)}\| = \|\phi(h^d(t+dt)) - \phi(h^{(d)}(t)\|$). In the context of random feedforward networks, we can consider piecewise linear activation functions as restrictions of $dh^{(d)}$ to a particular support set which is statistically dependent on $h^{(d)}$. This approach was developed by [23]. The key to our proof is providing a more direct and more generally applicable way of accounting for this dependence than originally provided by [23]. Specifically, our approach lets us derive the following, more general result, from which Corollaries 2.2, 2.3, and 2.4 follow easily.

Theorem 2.5 (**Trajectory growth in deep random sparse networks**). *Let $f_{NN}(x;\alpha,\mathcal{P},\mathcal{Q})$ be a random sparse network as defined in Section 1.1, with layers of width k. Let $\mathcal{P}$ and $\mathcal{Q}$ be such that the joint distribution over a vector of independent elements from both distributions is even. If $\mathbb{E}[|\boldsymbol{u}^\top w_{\mathcal{P}_i}|] \geq M\|\boldsymbol{u}\|$ for any constant vector $\boldsymbol{u}$, for all i, then*

$$\mathbb{E}[l(z^{(d)}(t))] \geq \left(\frac{\alpha M\sqrt{k}}{2}\right)^d \cdot l(x(t)), \tag{2.4}$$

for $x(t)$ a 1-dimensional trajectory in input space.

Remark 2.6. It is trivial to amend this result for networks where the width, distribution, and sparsity varies layer by layer, in which case the lower bound (2.4) is replaced by

$$\prod_{j=i}^{d}\left(\frac{\alpha_j M_j\sqrt{k_j}}{2}\right)\cdot l(x(t)).$$

Moreover, the bounds from Theorem 2.5 and Corollaries 2.2 - 2.4 hold true in the 0 bias case as well.

3. Proof of Theorem 2.5.

We prove Theorem 2.5 in three stages: i) We turn the problem into one of bounding from below the change in the length of an infinitesimal line segment; ii) we account simply and explicitly for the dependence generated by the ReLU activation; and iii) we break this dependence by taking advantage of the symmetry characterising this class of distributions. Supporting lemmas can be found in Section 4.

Proof.

Stage 1: For the first stage of proof, we will closely follow [23]. We are interested in deriving a lower bound of the form,

$$\mathbb{E}\left[\int_t \left|\left|\frac{dz^{(d)}(t)}{dt}\right|\right| dt\right] \geq C \cdot \int_t \left|\left|\frac{dx(t)}{dt}\right|\right| dt, \tag{3.1}$$

for some constant C. As noted by [23], it suffices to instead derive a bound of the form

$$\mathbb{E}\left[\|dz^{(d)}(t)\|\right] \geq C\|dx(t)\|,$$

since integrating over t yields the desired form. Our approach will be to derive a recurrence relation between $\|dz^{(d+1)}\|$ and $\|dz^{(d)}\|$, where we refrain from explicitly including the dependence of dz on t, for notational clarity.

Next, like [23], our proof relies on the observation that

$$\begin{aligned} dz^{(d+1)} &= \phi(W^{(d)}z^{(d)}(t+\delta t) + b^{(d)}) - \phi(W^{(d)}z^{(d)}(t) + b^{(d)}) \\ &= \phi^{(d)}(t+\delta t) - \phi^{(d)}(t) \\ &= d\phi^{(d)}, \end{aligned}$$

and that since ϕ is the ReLU operator, $\frac{d\phi}{dh_j^{(d)}}$ is either 0 or 1. When $z^{(d)}$ is fixed independently of $W^{(d)}$ and $b^{(d)}$, then $P(h_j^{(d)} = 0) = 0$ (see the preamble to Lemma 4.6 for more detail on this), and thus we need only note that $d\phi_j^{(d)} = dh_j^{(d)}$ when $h_j^{(d)} > 0$, and $d\phi_j^{(d)} = 0$ when $h_j^{(d)} < 0$. We define $\mathcal{A}^{(d)}$ to be the set of 'active nodes' in layer d; specifically,

$$\mathcal{A}^{(d)} := \{j : h_j^{(d)} > 0\},$$

and $I_{\mathcal{A}^{(d)}} \in \mathbb{R}^{k\times k}$ is defined as the matrix with ones on the diagonal entries indexed by set $\mathcal{A}^{(d)}$, and 0 everywhere else. We can then write

$$\begin{aligned} \|dz^{(d+1)}\| &= \|I_{\mathcal{A}^{(d)}}(h^{(d)}(t+dt) - h^{(d)}(t))\| \\ &= \|I_{\mathcal{A}^{(d)}}W^{(d)}dz^{(d)}\|. \end{aligned}$$

From here we will drop the weight index (d) to minimise clutter in the exposition.

It is at this point where we depart from the proof strategy used by [23]. The next steps in their proof depend heavily on the weight matrices in the network being Gaussian. For example, they require that a weight matrix after rotation has the same, i.i.d. distribution as the matrix before rotation. Instead, our proof can tackle a number of other, non-rotationally-invariant distributions, as well as sparse networks.

Stage 2: The next stage of the proof begins by noting that after conditioning on size of the set $\mathcal{A}$,

$$\mathbb{E}[\|I_{\mathcal{A}}Wdz^{(d)}\| \mid |\mathcal{A}|] = \mathbb{E}[\|\hat{W}dz^{(d)}\| \mid \hat{w}_i^\top z^{(d)} + \hat{b}_i > 0 \;\forall i, |\mathcal{A}|], \tag{3.2}$$

where $\hat{W} \in \mathbb{R}^{|\mathcal{A}| \times k}$ is the matrix comprised of the rows of W indexed by $\mathcal{A}$, and we denote the i-th row of $\hat{W}$ as $\hat{w}_i$, and the i-th entry of $\hat{b}$ as $\hat{b}_i$. Equation 3.2 follows since the elements of $Wdz^{(d)}$ are i.i.d., and $\mathcal{A}^{(d)}$ selects all entries whose corresponding entries in $h^{(d)}$ have positive values. Thus, in expectation, pre-multiplying by the matrix $I_{\mathcal{A}^{(d)}}$ is equivalent to considering $\hat{W}dz^{(d)}$ instead of $I_{\mathcal{A}}Wdz^{(d)}$ together with conditioning on the fact that every element in the vector $\hat{W}z^{(d)} + \hat{b}$ is positive.

This gives us

$$\mathbb{E}[\|I_{\mathcal{A}}Wdz^{(d)}\|\ ] = \mathbb{E}\left[\mathop{\mathbb{E}}_{\hat{w}_1}\mathop{\mathbb{E}}_{\hat{w}_2}\cdots\mathop{\mathbb{E}}_{\hat{w}_{|\mathcal{A}|}}\left[\sqrt{\sum_{i=1}^{|\mathcal{A}|}(\hat{w}_i^\top dz^{(d)})^2}\ \middle|\ \hat{w}_i^\top z^{(d)} + \hat{b}_i > 0\ \forall i, |\mathcal{A}|\right]\right] \tag{3.3}$$

$$= \mathbb{E}\left[\mathop{\mathbb{E}}_{\hat{w}_1}\mathop{\mathbb{E}}_{\hat{w}_2}\cdots\mathop{\mathbb{E}}_{\hat{w}_{|\mathcal{A}|}}\left[\sqrt{\sum_{i=1}^{|\mathcal{A}|}|\hat{w}_i^\top dz^{(d)}|^2}\ \middle|\ \hat{w}_i^\top z^{(d)} + \hat{b}_i > 0\ \forall i, |\mathcal{A}|\right]\right] \tag{3.4}$$

$$\geq \mathbb{E}\left[\sqrt{\sum_{i=1}^{|\mathcal{A}|}\mathop{\mathbb{E}}_{\hat{w}_i}[|\hat{w}_i^\top dz^{(d)}|\ |\hat{w}_i^\top z^{(d)} + \hat{b}_i > 0]^2}\right], \tag{3.5}$$

where (3.3) follows from the analysis above and the independence of each $\hat{w}_i$, (3.4) is trivial, and (3.5) follows from iteratively applying Jensen's inequality, after noting that $f(x) = \sqrt{x^2 + C}$ is convex for $x, C \geq 0$.

Now let J_i denote the (random) index set of the $\mathcal{P}$-distributed entries of $\hat{w}_i$, and let $w_{J_i}, dz_{J_i}^{(d)}, z_{J_i}^{(d)}$ denote the restrictions to the indices in J_i of $\hat{w}_i$, $dz^{(d)}$ and $z^{(d)}$ respectively. Then $\hat{w}_i^\top z^{(d)} = w_{J_i}^\top z_{J_i}^{(d)}$, and $\hat{w}_i^\top dz^{(d)} = w_{J_i}^\top dz_{J_i}^{(d)}$, such that, after conditioning on J_i, we have that

$$\mathbb{E}[\|\hat{W}p\|\ |\ \hat{w}_i^\top z^{(d)} + \hat{b}_i > 0\ \forall i, |\mathcal{A}|] \geq \underbrace{\mathbb{E}\left[\sqrt{\sum_{i=1}^{|\mathcal{A}|}\overbrace{\mathop{\mathbb{E}}_{J_i}\Big[\underbrace{\mathop{\mathbb{E}}_{w_{J_i}}[|w_{J_i}^\top dz_{J_i}^{(d)}|\ |w_{J_i}^\top z_{J_i}^{(d)} + \hat{b}_i > 0, J_i]}_{(*)}\Big]}^{(**)}{}^2}\right]}_{(***)}. \tag{3.6}$$

Stage 3: The third stage of the proof is to work our way from the inside out, lower bounding $(*)$ first, then $(**)$, and finally $(***)$.

Consider the expectation in $(*)$. Having conditioned on J_i, we can define $X = w_{J_i}^\top dz_{J_i}^{(d)}$ and $Y = w_{J_i}^\top z_{J_i}^{(d)} + \hat{b}_i$, such that lower bounding $(*)$ means lower bounding

$$\mathbb{E}[|X|\ |Y > 0]. \tag{3.7}$$

By assumption the joint distribution over $G = [w_{J_i,1}, \ldots, w_{J_i,k}, \hat{b}_i]^\top$ is even. The vector $H = [X, Y, w_{J_i,3} \ldots, w_{J_i,k}, \hat{b}_i]^\top$ is obtained by a linear transformation of G (which is invertible since $z^{(d)}$ is not parallel to $dz^{(d)}$). Thus by Lemma 4.1 (continuous) or Lemma 4.2 (discrete) this joint distribution over H is also even, and by Lemma 4.3 (continuous) or Lemma 4.4 (discrete), the joint distribution of $[X, Y]^\top$ is even too. We can therefore apply Lemma 4.5 (continuous) or Lemma 4.6 (discrete) and need only consider $\mathbb{E}[|X|]$, which is bounded as

$$\mathbb{E}[|X|] \geq M\|dz_{J_i}^{(d)}\|, \tag{3.8}$$

again by assumption.

Having bounded $(*)$, we average over J_i to get $(**)$, for which we can apply Lemma 4.7 to get

$$\mathbb{E}_{J_i}[M\|dz_{J_i}^{(d)}\|] \geq \alpha M\|dz^{(d)}\|. \tag{3.9}$$

Finally, we can bound $(***)$ as follows

$$\mathbb{E}[\|I_{\mathcal{A}}Wdz^{(d)}\|] \geq \mathbb{E}_{|\mathcal{A}|}\left[\sqrt{\sum_{i=1}^{|\mathcal{A}|}\alpha^2M^2\|dz^{(d)}\|^2}\right] \tag{3.10}$$

$$= \mathbb{E}_{|\mathcal{A}|}\left[\sqrt{|\mathcal{A}|\cdot\alpha^2M^2\|dz^{(d)}\|^2}\right] \tag{3.11}$$

$$\geq \mathbb{E}_{|\mathcal{A}|}\left[\frac{1}{\sqrt{k}\alpha M\|dz^{(d)}\|}\cdot|\mathcal{A}|\cdot\alpha^2M^2\|dz^{(d)}\|^2\right] \tag{3.12}$$

$$= \frac{\alpha M\|dz^{(d)}\|}{\sqrt{k}}\cdot\mathbb{E}[|\mathcal{A}|]. \tag{3.13}$$

where (3.10) is obtained by substituting the bound for $(**)$ into the inequality in (3.6), (3.11) follows since there is no dependence on i in the summed terms, and (3.12) follows since for any $0 \leq \gamma \leq \max(\gamma)$, $\sqrt{\gamma} \geq \frac{1}{\sqrt{max(\gamma)}}\gamma$, and $|\mathcal{A}|$ is at most k.

The proof is concluded by calculating $\mathbb{E}[|\mathcal{A}|]$. Since $|\mathcal{A}|$ is the number of entries in the vector $h^{(d)}$ which are positive, and each entry in that vector is an independent, centred random variable, $|\mathcal{A}|$ has a binomial distribution with probability $1/2$, and therefore an expected value of $k/2$. Plugging this in yields the final recursive relation between $\|dz^{(d+1)}\|$ and $\|dz^{(d)}\|$,

$$\mathbb{E}[\|dz^{(d+1)}\|] \geq \frac{\alpha M\sqrt{k}}{2}\|dz^{(d)}\|. \qquad \blacksquare$$

Iterative application of this result starting at the first layer yields the final result.

Let us illustrate the ease with which Corollaries 2.2, 2.3 and 2.4 are obtained. In the case of each distribution, we need to do two things. First, we must verify that the necessary assumption holds in the case of those distributions $\mathcal{P}$ and $\mathcal{Q}$: that the joint distribution over a vector of independent elements from both distributions is even. Second, we must derive a bound of the form $\mathbb{E}[|\boldsymbol{u}^\top\boldsymbol{w}|] \geq M\|\boldsymbol{u}\|$, where $w_i \sim \mathcal{P}$, and substitute M into Theorem 2.5.

When $\mathcal{P}$ and $\mathcal{Q}$ are centred Gaussians, the joint distribution over elements from one or both distributions is a multivariate Gaussian, with an even joint probability density function. Moreover, for $U = \boldsymbol{u}^\top \boldsymbol{w}$, $\mathbb{E}[|U|]$ has a closed form solution,

$$\mathbb{E}[|U|] = \frac{\sqrt{2}\sigma_w}{\sqrt{\pi}} \|\boldsymbol{u}\|.$$

When $\mathcal{P}$ and $\mathcal{Q}$ are centred uniform distributions, the joint distribution is uniform over the polygon bounded in each dimension by the symmetric bounds $[-C_w, C_w]$ or $[-C_b, C_b]$, and thus is even. Next, to bound $\mathbb{E}[|U|]$, we apply the Marcinkiewicz-Zygmund inequality with $p = 1$, using the optimal A_1 from Lemmas 4.8 and 4.9, to get that

$$\mathbb{E}[|U|] \geq \frac{C_w}{2\sqrt{2}} \|\boldsymbol{u}\|;$$

for details of this derivation, see Lemma 4.10.

Likewise, when $\mathcal{P}$ and $\mathcal{Q}$ are uniform distributions over discrete, symmetric, finite sets $\mathcal{W}$ and $\mathcal{B}$ respectively, we make a discrete analogue of the argument made in the continuous uniform case to confirm the necessary assumption holds. Bounding $\mathbb{E}[|U|]$ in this case also follows from a very similar argument to that made in the continuous case, detailed in full in Lemma 4.11, yielding

$$\mathbb{E}[|U|] \geq \frac{\sum_{w \in \mathcal{W}} |w|}{\sqrt{2} N_w} \|\boldsymbol{u}\|.$$

4. Lemmas used in proof of Theorem 2.5.

Lemma 4.1. *Let $f_X(\mathbf{x})$ be an even joint probability density function over random vector $X \in \mathbb{R}^k$. Let $A \in \mathbb{R}^{k \times k}$ be an invertible linear transformation such that $Y = AX$. Then the joint density $f_Y(\mathbf{y})$ is also even.*

Proof. Wlog we assume f_X is defined on $\mathbb{R}^k$. To calculate the density over $Y \in \mathbb{R}^k$ we make a change of variables such that

$$f_Y(\boldsymbol{y}) = f_X(A^{-1}\boldsymbol{y})|A^{-1}|. \tag{4.1}$$

Since A is one-to-one, we have that $f_X(\boldsymbol{x}) = f_X(A^{-1}\boldsymbol{y})$ for some $\boldsymbol{y}$, and f_X is even, so $f_X(A^{-1}\boldsymbol{y}) = f_X(-(A^{-1}\boldsymbol{y})) = f_X(A^{-1}(-\boldsymbol{y}))$ for all $\boldsymbol{y}$. Putting this together completes the proof,

$$f_Y(\boldsymbol{y}) = f_X(A^{-1}\boldsymbol{y})|A^{-1}| = f_X(A^{-1}(-\boldsymbol{y}))|A^{-1}| = f_Y(-\boldsymbol{y}). \tag{4.2}$$

∎

Lemma 4.2. *Let $f_X(\mathbf{x})$ be an even joint probability mass function over random vector $X \in \mathbb{R}^k$. Let $A \in \mathbb{R}^{k \times k}$ be an invertible linear transformation such that $Y = AX$. Then the joint mass function $f_Y(\boldsymbol{y})$ is also even.*

Proof. f_X is defined on some discrete, finite, symmetric set $\mathcal{X}$. To calculate the density over $Y \in \mathcal{Y} := \{A\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$ we make a change of variables such that

$$f_Y(\mathbf{y}) = \sum_{\boldsymbol{x} \in \{A\boldsymbol{x} = \boldsymbol{y}\}} f_X(\mathbf{x}). \tag{4.3}$$

Since A is one-to-one, we have that $f_X(\boldsymbol{x}) = f_X(A^{-1}\boldsymbol{y})$ for some $\boldsymbol{y}$, and f_X is even, so $f_X(A^{-1}\boldsymbol{y}) = f_X(-(A^{-1}\boldsymbol{y})) = f_X(A^{-1}(-\boldsymbol{y}))$ for all $\boldsymbol{y}$. Putting this together completes the proof,

$$f_Y(\mathbf{y}) = \sum_{\boldsymbol{x}\in\{A\boldsymbol{x}=\boldsymbol{y}\}} f_X(A^{-1}\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\{A\boldsymbol{x}=\boldsymbol{y}\}} f_X(A^{-1}(-\boldsymbol{y})) = f_Y(-\mathbf{y}). \tag{4.4}$$ ■

Lemma 4.3. *Let $f_{X_1,\dots,X_k}(x_1,\dots,x_k)$ be an even probability density function. Then $f_{X_1,\dots,X_{k-1}}(x_1,\dots,x_{k-1}) = \int_{-\infty}^{\infty} f_{X_1,\dots,X_k}(x_1,\dots,x_k)dx_k$ is also even.*

Proof.

$$f_{X_1,\dots,X_{k-1}}(x_1,\dots,x_{k-1}) = \int_{-\infty}^{\infty} f_{X_1,\dots,X_k}(x_1,\dots,x_k)dx_k \tag{4.5}$$

$$= \int_{-\infty}^{\infty} f_{X_1,\dots,X_k}(-x_1,\dots,-x_k)dx_k \tag{4.6}$$

$$= \int_{-\infty}^{\infty} f_{X_1,\dots,X_k}(-x_1,\dots,-x_{k-1},x_k)dx_k \tag{4.7}$$

$$= f_{X_1,\dots,X_{k-1}}(-x_1,\dots,-x_{k-1}). \tag{4.8}$$

Equalities 4.5 and 4.8 follow from the definition of marginalisation of random variables. Line 4.6 follows from the assumption that $f_{X_1,\dots,X_k}$ is even, and the line 4.7 follows from the change of variables: $-x_k \longrightarrow x_k$. ■

Lemma 4.4. *Let $X_1,\dots,X_k$ be discrete random variables with symmetric support sets $\mathcal{X}_1,\dots,\mathcal{X}_k$ respectively, i.e. $x_i \in \mathcal{X}_j \iff -x_i \in \mathcal{X}_j$. Let $P(X_1 = x_1,\dots,X_k = x_k)$ be an even probability mass function such that $P(X_1 = x_1,\dots,X_k = x_k) = P(X_1 = -x_1,\dots,X_k = -x_k)$. Then $P(X_1 = x_1,\dots,X_{k-1} = x_{k-1})$ is also even.*

Proof.

$$P(X_1 = x_1,\dots,X_{k-1} = x_{k-1}) = \sum_{x_k\in\mathcal{X}_k} P(X_1 = x_1,\dots,X_k = x_k) \tag{4.9}$$

$$= \sum_{x_k\in\mathcal{X}_k} P(X_1 = -x_1,\dots,X_k = -x_k) \tag{4.10}$$

$$= \sum_{-x_k\in\mathcal{X}_k} P(X_1 = -x_1,\dots,X_k = x_k) \tag{4.11}$$

$$= \sum_{x_k\in\mathcal{X}_k} P(X_1 = -x_1,\dots,X_k = x_k) \tag{4.12}$$

$$= P(X_1 = -x_1,\dots,X_{k-1} = -x_{k-1}). \tag{4.13}$$ ■

Lines 4.9 and 4.13 follow from the definition of marginal distributions, (4.10) follows by assumption, (4.11) follows fro a change of variables, and (4.12) follows since summing over $-x_k$ is equivalent to summing over x_k.

Lemma 4.5. *Let X and Y be random variables with an even joint probability density function $f_{XY}(x, y)$. Then*

$$\mathbb{E}[|X| \mid Y > 0] = \mathbb{E}[|X|]$$

Proof. Letting $|X| = Z$, we can make a straightforward change of variables to calculate the joint distribution $f_{ZY}(z, y)$, which works out to be

$$f_{ZY}(z, y) = f_{XY}(z, y) + f_{XY}(-z, y)$$

for $z \geq 0$ and $y \in \mathbb{R}$. Then we have that

$$\begin{aligned}
\mathbb{E}[Z|Y > 0] &= \int_0^\infty z \cdot f_{Z|Y>0}(z|y > 0)dz \\
&= \int_0^\infty z \cdot \frac{f_{Z,Y>0}(z, y > 0)}{\int_0^\infty f_Y(y)dy}dz \\
&= 2\int_0^\infty z \cdot f_{Z,Y>0}(z, y > 0)dz \\
&= 2\int_0^\infty z \int_0^\infty f_{ZY}(z, y)dydz \\
&= 2\int_0^\infty z \int_0^\infty (f_{XY}(z, y) + f_{XY}(-z, y))dydz.
\end{aligned}$$

One the other hand, we have that

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^\infty z \cdot f_Z(z)dz \\
&= \int_0^\infty z \cdot (f_X(z) + f_X(-z))dz \\
&= 2\int_0^\infty z \cdot f_X(z)dz \\
&= 2\int_0^\infty z \cdot \int_{-\infty}^\infty f_{XY}(z, y)dydz \\
&= 2\int_0^\infty z \cdot \left(\int_{-\infty}^0 f_{XY}(z, y)dy + \int_0^\infty f_{XY}(z, y)dy\right) dz.
\end{aligned}$$

Comparing the expressions for $\mathbb{E}[Z|Y > 0]$ and $\mathbb{E}[Z]$, we can see that they are equal if

$$\int_{-\infty}^0 f_{XY}(z, y)dy = \int_0^\infty f_{XY}(-z, y)dy.$$

A change of variables on the left hand side from y to $-y$ yields

$$\int_{-\infty}^0 f_{XY}(z, y)dy = \int_0^\infty f_{XY}(z, -y)dy. \qquad \blacksquare$$

and by assumption, we know that $f_{XY}(z, -y) = f_{XY}(-z, y)$ since f_{XY} is even, which completes the proof.

Lemma 4.5 implicitly makes use of the fact that $P(Y=0)=0$, which follows from w_{J_i} and $\hat{b}_i$ being continuous random variables, and $Y = w_{J_i}^\top z_{J_i} + \hat{b}_i$, with z_{J_i} being fixed independent of w_{J_i}. We similarly make use of the fact that $P(Y=0)=0$ in the application of Lemma 4.6, though that this is true is less immediately apparent in the discrete case. For clarity, let us define $\boldsymbol{v} := [w_{J_i}, \hat{b}_i]$, the concatenation of w_{J_i} and $\hat{b}_i$, and $\hat{\boldsymbol{z}} := [z_{J_i}, 1]$, the concatenation of z_{J_i} and 1, such that $Y = \boldsymbol{v}^\top \hat{\boldsymbol{z}}$. Associated with the discrete distribution over $\boldsymbol{v}$ there are $N_w^{|J_i|} N_b$ possible discrete random vectors in $\mathbb{R}^{|J_i|+1}$. The set of vectors $\hat{\boldsymbol{z}} \in \mathbb{R}^{|J_i|+1}$ orthogonal to such a discrete set is measure zero, and as such for $\hat{\boldsymbol{z}}$ fixed independent of the choice of the discrete measure $\boldsymbol{v}$ we have $P(\boldsymbol{v}^\top \hat{\boldsymbol{z}} = 0) = 0$. If however $\hat{\boldsymbol{z}}$ were selected with knowledge of the discrete distribution $\boldsymbol{v}$ then one of two cases will occur; either $\boldsymbol{v}^\top \hat{\boldsymbol{z}} \neq 0$, or $\hat{\boldsymbol{z}}$ is selected to be from the measure zero set of vectors orthogonal to any of the $N_w^{|J_i|} N_b$ vectors generated by $\boldsymbol{v}$. In the latter case, the assumptions in Lemma 4.6 of $\mathcal{Y}$ excluding 0 would not be satisfied. In such an adversarial case there would be a discrepancy between $\mathbb{E}[|X| \mid Y>0]$ and $\mathbb{E}[|X|]$ which would shrink as the proportion of the $N_w^{|J_i|} N_b$ vectors generated by $\boldsymbol{v}$ to which that particular $\hat{\boldsymbol{z}}$ is orthogonal.

Lemma 4.6. *Let X and Y be discrete random variables with finite, symmetric support sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, where $0 \notin \mathcal{Y}$, and an even joint probability mass function $f_{XY}(x,y)$ such that $P(X=x, Y=y) = P(X=-x, Y=-y)$. Then*

$$\mathbb{E}[|X| \mid Y > 0] = \mathbb{E}[|X|]$$

Proof. Letting $|X| = Z$, we can make a change of variables to obtain the joint mass function $f_{ZY}(z,y)$, which works out to be

$$f_{ZY}(z,y) = \begin{cases} f_{XY}(z,y) + f_{XY}(-z,y) & \text{for } \ (z,y) \text{ where } z \in \mathcal{X}^+ \text{ and } y \in \mathcal{Y} \\ f_{XY}(z,y) & \text{for } (z,y) \text{ where } z = 0 \text{ and } \in \mathcal{Y} \end{cases}$$

where $\mathcal{X}^+$ is the set of all positive elements of $\mathcal{X}$.

Next, we have that

$$\begin{aligned}
\mathbb{E}[Z|Y>0] &= \sum_{z \in \mathcal{X}^+} zP(Z=z|Y>0) \\
&= \sum_{z \in \mathcal{X}^+} z \frac{P(Z = z \cap Y > 0)}{P(Y>0)} && (4.14) \\
&= 2 \sum_{z \in \mathcal{X}^+} zP(Z=z \cap Y>0) && (4.15) \\
&= 2 \sum_{z \in \mathcal{X}^+} \sum_{y \in \mathcal{Y}^+} zP(Z = z \cap Y = y) \\
&= 2 \sum_{z \in \mathcal{X}^+} \sum_{y \in \mathcal{Y}^+} z \left(f_{XY}(z,y) + f_{XY}(-z,y) \right). && (4.16)
\end{aligned}$$

On the other hand, we have

$$\mathbb{E}[Z] = \sum_{z\in\mathcal{X}^+} zP(Z=z) \tag{4.17}$$
$$= \sum_{z\in\mathcal{X}^+} z\left(f_X(z)+f_X(-z)\right) \tag{4.18}$$
$$= 2\sum_{z\in\mathcal{X}^+} zf_X(z) \tag{4.19}$$
$$= 2\sum_{z\in\mathcal{X}^+}\sum_{y\in\mathcal{Y}} zf_{XY}(z,y) \tag{4.20}$$
$$= 2\sum_{z\in\mathcal{X}^+}\left(\sum_{y\in\mathcal{Y}^+} zf_{XY}(z,y) + \sum_{y\in\mathcal{Y}^-} zf_{XY}(z,y)\right). \tag{4.21}$$

Next, we not that

$$\sum_{y\in\mathcal{Y}^-} zf_{XY}(z,y) = \sum_{y\in\mathcal{Y}^+} zf_{XY}(z,-y)$$
$$= \sum_{y\in\mathcal{Y}^+} zf_{XY}(-z,y).$$ ■

Thus the expressions in 4.16 and 4.21 are equal, which completes the proof.

Lemma 4.7 (Expected norm of a random sub-vector). *Let $\boldsymbol{u}\in\mathbb{R}^k$ be a fixed vector and let $J\subseteq\{1,2,\ldots,k\}$ be a random index set, where the probability of any index from 1 to k appearing in any given sample is independent and equal to α. Then, defining $\boldsymbol{u}_J$ to be the vector comprised only of the elements of $\boldsymbol{u}$ indexed by J, we can lower bound the expectation of the norm of this subvector by*

$$\mathbb{E}_J[\|\boldsymbol{u}_J\|] \geq \alpha\|\boldsymbol{u}\|. \tag{4.22}$$

Proof. First, we bound the expectation of the norm in terms of the expectation of the squared norm as follows:

$$\mathbb{E}[\|\boldsymbol{u}_J\|] = \mathbb{E}[\sqrt{\sum_{j\in J} u_{J,j}^2}] \tag{4.23}$$
$$\geq \frac{1}{\|\boldsymbol{u}\|}\mathbb{E}[\sum_{j\in J} u_{J,j}^2]. \tag{4.24}$$

This follows because for any $0\leq\gamma\leq\max(\gamma)$, $\sqrt{\gamma}\geq\frac{1}{\sqrt{max(\gamma)}}\gamma$.

Next we note that $\sum_{j\in J} u_{J,j}^2$ is exactly equivalent to $\sum_{i=1}^k u_i^2 B_i$, a weighted sum of k iid Bernoulli random variables B_i with $p=\alpha$, and so

$$\mathbb{E}[\sum_{j\in J} u_{J,j}^2] = \sum_{i=1}^k u_i^2\cdot\mathbb{E}[B] \tag{4.25}$$
$$= \|\boldsymbol{u}\|^2\cdot\alpha. \tag{4.26}$$

Substituting this into inequality 4.24 completes the proof,

$$\mathbb{E}[\|\boldsymbol{u}_J\|] \geq \alpha\|\boldsymbol{u}\|. \qquad \blacksquare$$

Lemmas 4.8 and 4.9 are taken from [6], and are restated here for completeness.

Lemma 4.8 (Marcinkiewicz-Zygmund Inequality ([6])). *Let $X_1, \ldots, X_n$ be $n \in \mathbb{N}$ independent and centered real random variables defined on some probability space (Ω, A, P) with $\mathbb{E}[|Xi|^p] < \infty$ for every $i \in \{1, ..., n\}$ and for some $p > 0$. Then for every $p \geq 1$ there exist positive constants A_p and B_p depending only on p such that*

$$A_p\mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{p/2}\right] \leq \mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right] \leq B_p\mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{p/2}\right]. \tag{4.27}$$

Lemma 4.9 (Optimal constants for Marcinkiewicz-Zygmund Inequality ([6])). *Let Γ denote the Gamma function and let p_0 be the solution of the equation $\Gamma(\frac{p+1}{2}) = \sqrt{\pi}/2$ in the interval $(1, 2)$, i.e. $p_0 \approx 1.84742$. Then for every $p > 0$ it holds:*

$$A_{p,opt} = \begin{cases} 2^{p/2-1}, & 0 < p \leq p_0 \\ 2^{p/2} \cdot \frac{\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}, & p_0 < p < 2 \\ 1 & 2 \leq p < \infty \end{cases} \tag{4.28}$$

and

$$B_{p,opt} = \begin{cases} 1 & 0 < p \leq 2 \\ 2^{p/2} \cdot \frac{\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}, & 2 < p < \infty \end{cases} \tag{4.29}$$

Lemma 4.10. *Let $X = \sum_i \alpha_i w_i$, where $w_i \sim \mathcal{U}(-C, C)$ where w_i are uniform random scalars. Then*

$$\mathbb{E}[|X|] \geq \frac{C}{2\sqrt{2}}\|\alpha\|.$$

Proof. Defining $X_i = \alpha_i w_i$, we can then apply the Marcinkiewicz-Zygmund inequality with $p = 1$, using the optimal A_1 from Lemma 4.9 to get that

$$\mathbb{E}[|X|] = \mathbb{E}\left[\left|\sum_{i=1}^k X_i\right|\right] \geq \frac{1}{\sqrt{2}}\mathbb{E}\left[\sqrt{\sum_{i=1}^k X_i^2}\right].$$

Next we use the same tricks as early in the proof of the general case:

$$\frac{1}{\sqrt{2}}\mathbb{E}\left[\sqrt{\sum_{i=1}^k X_i^2}\right] = \frac{1}{\sqrt{2}}\mathbb{E}\left[\sqrt{\sum_{i=1}^k |X_i|^2}\right] \tag{4.30}$$

$$\geq \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^k \mathbb{E}[|X_i|]^2}, \tag{4.31}$$

where line 4.30 is trivial and line 4.31 follows from a repeated application of Jensen's inequality.

To calculate $\mathbb{E}[|X_i|]$ we note that $X_i = \alpha_i w_i$ is uniformly distributed as $X_i \sim U(-|\alpha_i|C, |\alpha_i|C)$, and thus

$$\mathbb{E}[|X_i|] = \frac{C|\alpha_i|}{2},$$

and so

$$\begin{aligned}
\mathbb{E}[|X|] &\geq \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\mathbb{E}[|X_i|]^2} \\
&= \frac{1}{\sqrt{2}}\sqrt{\frac{C^2}{4}\sum_{i=1}^{k}|\alpha_i|^2} \\
&= \frac{C}{2\sqrt{2}}\|\alpha\|.
\end{aligned}$$

■

Lemma 4.11. *Let $X = \sum_i \alpha_i w_i$, where w_i are uniformly sampled from some discrete symmetric sample space $\mathcal{W}$. Then*

$$\mathbb{E}[|X|] \geq \frac{\sum_{w\in\mathcal{W}}|w|}{\sqrt{2}N_w}\|\alpha\|.$$

Proof. Defining $X_i = \alpha_i w_i$, we follow exactly the same steps as in the first part of the proof of Lemma 4.10, to get that

$$\mathbb{E}[|X|] \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\mathbb{E}[|X_i|]^2}.$$

To calculate $\mathbb{E}[|X_i|]$ we note that $X_i = \alpha_i w_i$ is uniformly sampled from $\alpha_i\mathcal{W}$ and thus

$$\mathbb{E}[|X_i|] = \frac{|\alpha_i|\sum_{w\in\mathcal{W}}|w|}{N_w},$$

and so

$$\begin{aligned}
\mathbb{E}[|X|] &\geq \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\mathbb{E}[|X_i|]^2} \\
&= \frac{1}{\sqrt{2}}\sqrt{\frac{(\sum_{w\in\mathcal{W}}|w|)^2}{N_w^2}\sum_{i=1}^{k}|\alpha_i|^2} \\
&= \frac{\sum_{w\in\mathcal{W}}|w|}{\sqrt{2}N_w}\|\alpha\|.
\end{aligned}$$

■

Lemma 4.12. *Let $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^k$ be discrete sets with finite cardinality, and $g : \mathcal{W} \longrightarrow \mathcal{X}$ be a one-to-one transformation. Then if $P(W = \mathbf{w}) = P(W_1 = w_1, \ldots, W_k = w_k) = C$ for all $\mathbf{w} \in \mathcal{W}$, where C is constant, then $P(X = \mathbf{x}) = C$ for all $\mathbf{x} \in \mathcal{X}$*

Proof.

$$P(X = \mathbf{x}) = \sum_{\mathbf{w} \in \{g(\mathbf{w}) = \mathbf{x}\}} P(W = \mathbf{w}) \tag{4.32}$$

$$= C. \tag{4.33}$$

Equation 4.32 is a change of variables, and (4.33) follows from the fact the there is only ever one term in the sum, since g is one-to-one. ■

5. Numerical Simulations.

In this section we demonstrate, through numerical simulations, how the relationships between the the network's distributional and architectural properties observed in practice compare with those described in the lower bounds of Corollaries 2.2 - 2.4. To this end, we use as our trajectory a straight line between two (normalised) MNIST datapoints[1], discretized into 10000 pieces. For each combination of distribution and parameters, we pass the aforementioned line through 100 different deep neural networks of width 784, and average the results. Specifically, we consider three different networks types, sparse-Gaussian, sparse-uniform, and sparse-discrete networks, from Definitions 1.2 - 1.4 respectively. For each distribution we consider different values of network fractional density α ranging from 0.1 to 1. In the sparse-Gaussian networks, non-zero weights are sampled from $\mathcal{N}(0, \sigma_w^2/k)$, and biases from $\mathcal{N}(0, 0.01^2)$. In the sparse-Uniform networks, non-zero weights are sampled from $\mathcal{U}(-C/\sqrt{k}, C/\sqrt{k})$, and biases from $\mathcal{U}(-0.01, 0.01)$. In the sparse-discrete networks, non-zero weights are uniformly sampled from $\mathcal{W} := (1/\sqrt{k}) \odot \{-C, -(C+1), \ldots, C-1, C\}$, and biases from $\mathcal{B} := \{-0.01, 0.01\}$. We do this for a variety of σ_w and C values. The results are shown in Figures 2 and 3.

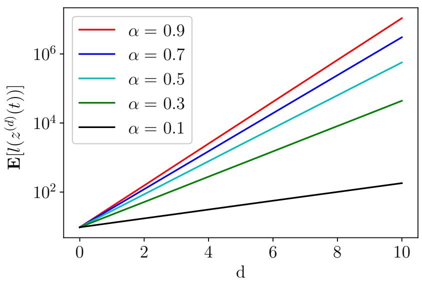
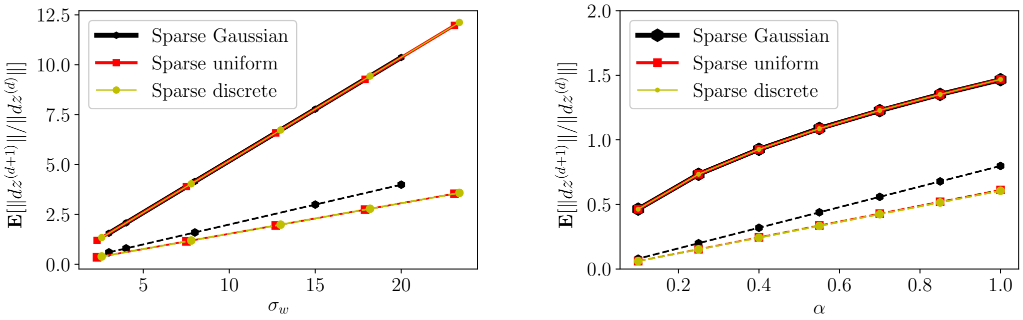


Figure 2: Expected length of a line connecting two MNIST data points as it passes through a sparse-Gaussian deep network, plotted at each layer d.

Figure 2 plots the average length of the trajectory at layer d of a sparse-Gaussian network, with $\sigma_w = 6$ and for different choices of sparsity ranging from 0.1 to 0.9. We see exponential

[1]In this experiment we chose the 101^{st} and 1001^{st} points from the MNIST test set, but the choice of points does not qualitatively change the results.

increase of expected length with depth even in sparse networks, with smaller slopes for smaller α (higher sparsity). In Figures 3a and 3b we plot the growth ratio of a small piece of the



(a) (b)

Figure 3: Expected growth factor, that is, the expected ratio of the length of any very small line segment in layer $d+1$ to its length in layer d. Figure 3a shows the dependence on the variance of the weights' distribution, and Figure 3b shows the dependence on sparsity.

trajectory from one layer to the next, averaged over all pieces, at all layers, and across all 100 networks for a given distribution. This $\mathbb{E}[\|dz^{(d+1)}\|/\|dz^{(d)}\|]$ corresponds to the base of the exponential in our lower bound. The solid lines reflect the observed averages of this ratio, while the dashed lines reflect the lower bound from Corollaries 2.2, 2.3, and 2.4. Figure 3a illustrates the dependence on the standard deviation of the respective distributions (before scaling by $1/\sqrt{k}$), with α fixed at $\alpha = 0.5$. We observe both that the lower bounds clearly hold, and that the dependence on σ_w is linear in practice, exactly as we expect from our lower bounds. Figure 3b shows the dependence of this ratio on the sparsity parameter α, where we have fixed $\sigma_w = 2$ for all distributions. Once again, the lower bounds hold, but in this case the observed α dependence is not exactly linear, but rather appears closer to $\sqrt{\alpha}$. The likely source of this qualitative discrepancy is the use of Lemma 4.7, to lower bound

$$\mathbb{E}_{J_i}[\|dz_{J_i}\|] \geq \alpha\|dz\|, \tag{5.1}$$

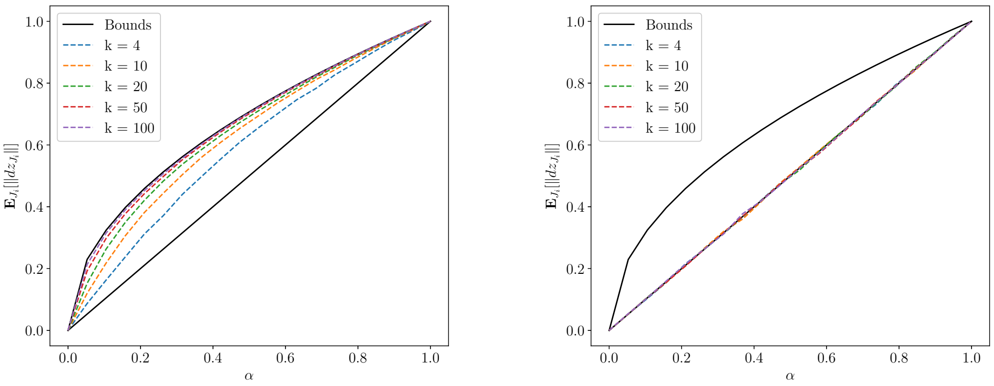
used in (3.9) in Stage 3 of the proof of Theorem2.5. It is straightforward to derive an *upper* bound for this same quantity, as

$$\mathbb{E}_{J_i}[\|dz_{J_i}\|] \leq \sqrt{\alpha}\|dz\|, \tag{5.2}$$

first using Jensen's inequality to get that $\mathbb{E}_{J_i}[\sqrt{\|dz_{J_i}\|^2}] \leq \sqrt{\mathbb{E}[\|dz_{J_i}\|^2]}$, and then using the strategy from the proof of Lemma 4.7 to get $\mathbb{E}[\|dz_{J_i}\|^2] = \alpha\|dz\|^2$.

To explore this discrepancy between the observed growth ratio and the lower and upper bounds from (5.1) and (5.2), we consider different fixed vectors $dz \in \mathbb{R}^k$, and average over subvectors dz_{J_i}. Specifically, we calculated the expected value of a subvector dz_{J_i} containing only the entries of dz indexed by J_i, where $J_i \subseteq \{1, 2, \ldots, k\}$ is a random index set, where the probability of any index from 1 to k appearing in any given sample is independent and equal to α. Figure 4a shows the results when dz a realisation of the uniform distribution over the unit sphere, with different dimensions k.

For even moderately large k, and vectors dz where most entries are roughly this same magnitude, this upper bound is very tight, such that the expected norm of the subvector generally behaves like $\sqrt{\alpha}\|dz\|$, not $\alpha\|dz\|$. However, it is also possible to construct an example



(a) (b)

Figure 4: The dependence on α and k of expected value of a subvector dz_{J_i}. In Figure 4a, dz is a realisation of the uniform distribution over the unit sphere. In Figure 4b, dz has the first entry equal to 1, and the rest zeros.

where the lower bound is tight, by letting dz have only a single non-zero entry, which case $\mathbb{E}[\|\boldsymbol{u}_J\|] = \alpha\|\boldsymbol{u}\|$ (see Figure 4b). While the former case, with entries of dz mostly of the same order, is typical, especially past the first few layers of the network, the bound cannot be improved without further assumptions on $\|dz\|$.

One striking observation in Figures 3a and 3b is that for a given σ_w, the observed $\mathbb{E}[\|dz^{(d+1)}\|/\|dz^{(d)}\|]$ matches perfectly across all three distributions, for different values of σ_w and different α. This remains true when we repeat the experiments with different datapoints, and with points chosen uniformly at random in a high-dimensional space, both when the trajectory considered is a straight line and when it is not (e.g. arcs in two or more dimensions), and the resulting figures are visually indistinguishable from Figures 3a and 3b. Another implication of these experiments is that they give some guidance for how to trade off weight scale against sparsity depending on the desired network properties. For example, Figure 3b considers the initialisation scheme with $\sigma_w = 2/\sqrt{k}$. We see that the empirically observed growth factor from one layer to the next is approximately 1.5 when the matrices are dense ($\alpha = 1$), while the growth factor is 1 with $\alpha \approx 0.5$, and less than one as α decreases further.

6. Conclusion. Our proof strategy and results generalise and extend previous work by [23] to develop theoretical guarantees lower bounding expected trajectory growth through deep neural networks for a broader class of network weight distributions and the setting of sparse networks. We illustrate this approach with Gaussian, uniform, and discrete valued random weight matrices with any sparsity level.

Acknowledgments. We would like to thank Maithra Raghu for her kind response to early enquiries about her paper [23] which inspired this work.

REFERENCES

[1] A. Aghasi, A. Abdi, N. Nguyen, and J. Romberg, *Net-trim: Convex pruning of deep neural networks with performance guarantee*, in Advances in Neural Information Processing Systems, 2017, pp. 3177–3186.

[2] S. Ahmad and L. Scheinkman, *How can we be so dense? the benefits of using highly sparse representations*, arXiv preprint arXiv:1903.11257, (2019).

[3] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, *Optimal approximation with sparsely connected deep neural networks*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 8–45.

[4] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals, and Systems (MCSS), 5 (1992), pp. 455–455.

[5] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, *Empirical study of the topology and geometry of deep networks*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3762–3770.

[6] D. Ferger, *Optimal constants in the marcinkiewicz–zygmund inequalities*, Statistics & Probability Letters, 84 (2014), pp. 96–101.

[7] J. Frankle and M. Carbin, *The lottery ticket hypothesis: Finding sparse, trainable neural networks*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=rJl-b3RcF7.

[8] R. Giryes, G. Sapiro, and A. M. Bronstein, *Deep neural networks with random gaussian weights: A universal classification strategy?*, IEEE Transactions on Signal Processing, 64 (2016), pp. 3444–3457.

[9] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[10] W. H. Guss and R. Salakhutdinov, *On characterizing the capacity of neural networks using algebraic topology*, arXiv preprint arXiv:1802.04443, (2018).

[11] S. Han, J. Pool, J. Tran, and W. Dally, *Learning both weights and connections for efficient neural network*, in Advances in neural information processing systems, 2015, pp. 1135–1143.

[12] B. Hanin and D. Rolnick, *Complexity of linear regions in deep networks*, in International Conference on Machine Learning, 2019, pp. 2596–2604.

[13] K. He, Y. Wang, and J. Hopcroft, *A powerful generative model using random weights for the deep image representation*, in Advances in Neural Information Processing Systems, 2016, pp. 631–639.

[14] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural networks, 2 (1989), pp. 359–366.

[15] W. Huang, P. Hand, R. Heckel, and V. Voroninski, *A provably convergent scheme for compressive sensing under random generative priors*, arXiv preprint arXiv:1812.04176, (2018).

[16] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, *Binarized neural networks*, in Advances in neural information processing systems, 2016, pp. 4107–4115.

[17] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, *Quantized neural networks: Training neural networks with low precision weights and activations*, The Journal of Machine Learning Research, 18 (2017), pp. 6869–6898.

[18] H. Li, S. De, Z. Xu, C. Studer, H. Samet, and T. Goldstein, *Training quantized nets: A deeper understanding*, in Advances in Neural Information Processing Systems, 2017, pp. 5811–5821.

[19] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, *The expressive power of neural networks: A view from the width*, in Advances in neural information processing systems, 2017, pp. 6231–6239.

[20] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, *Multi-layer generalized linear estimation*, in 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 2098–2102.

[21] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, *On the number of linear regions of deep neural networks*, in Advances in neural information processing systems, 2014, pp. 2924–2932.

[22] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos*, in Advances in Neural Information Processing Systems 29, 2016, pp. 3360–3368.

[23] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, *On the expressive power of deep neural networks*, in Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR.org, 2017, pp. 2847–2854.

[24] U. Shaham, A. Cloninger, and R. R. Coifman, *Provable approximation properties for deep neural networks*, Applied and Computational Harmonic Analysis, 44 (2018), pp. 537–557.

[25] S. Wu, G. Li, F. Chen, and L. Shi, *Training and inference with integers in deep neural networks*, in International Conference on Learning Representations, 2018, https://openreview.net/forum?id=HJGXzmspb.

[26] H. Yang, Y. Zhu, and J. Liu, *Energy-constrained compression for deep neural networks via weighted sparse projection and layer input masking*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=BylBr3C9K7.