On Outer Bi-Lipschitz Extensions of Linear Johnson-Lindenstrauss Embeddings of Low-Dimensional Submanifolds of $\mathbb{R}^N$

Mark A. Iwen* and Mark Philip Roach†

June 8, 2022

Abstract

Let $\mathcal{M}$ be a compact d-dimensional submanifold of $\mathbb{R}^N$ with reach τ and volume $V_{\mathcal{M}}$. Fix $\epsilon \in (0,1)$. In this paper we prove that a nonlinear function $f : \mathbb{R}^N \to \mathbb{R}^m$ exists with $m \leq C\left(d/\epsilon^2\right) \log\left(\frac{\sqrt[d]{V_{\mathcal{M}}}}{\tau}\right)$ such that

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|f(\mathbf{x})-f(\mathbf{y})\|_2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2$$

holds for all $\mathbf{x} \in \mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$. In effect, f not only serves as a bi-Lipschitz function from $\mathcal{M}$ into $\mathbb{R}^m$ with bi-Lipschitz constants close to one, but also approximately preserves all distances from points not in $\mathcal{M}$ to all points in $\mathcal{M}$ in its image. Furthermore, the proof is constructive and yields an algorithm which works well in practice. In particular, it is empirically demonstrated herein that such nonlinear functions allow for more accurate compressive nearest neighbor classification than standard linear Johnson-Lindenstrauss embeddings do in practice.

1 Introduction

The classical Kirszbraun theorem [15] ensures that a Lipschitz continuous function $f : S \to \mathbb{R}^m$ from a subset $S \subset \mathbb{R}^N$ into $\mathbb{R}^m$ can always be extended to a function $\tilde{f} : \mathbb{R}^N \to \mathbb{R}^m$ with the same Lipschitz constant as f. More recently, similar results have been proven for bi-Lipschitz functions, $f : S \to \mathbb{R}^m$, from $S \subset \mathbb{R}^N$ into $\mathbb{R}^m$ in the theoretical computer science literature. In particular, it was shown in [18] that outer extensions of such bi-Lipschitz functions f, $\tilde{f} : \mathbb{R}^N \to \mathbb{R}^{m+1}$, exist which both (i) approximately preserve f's Lipschitz constants, and which (ii) satisfy $\tilde{f}(\mathbf{x}) = (f(\mathbf{x}), 0)$ for all $\mathbf{x} \in S$. Narayanan and Nelson [19] then applied similar outer extension methods to a special class of the linear bi-Lipschitz maps guaranteed to exist for any given finite set $S \subset \mathbb{R}^N$ by Johnson-Lindenstrauss (JL) lemma [14] in order prove the following remarkable result: For each finite set $S \subset \mathbb{R}^N$ and $\epsilon \in (0,1)$ there exists a **terminal embedding of** S, $f : \mathbb{R}^N \to \mathbb{R}^{\mathcal{O}\left(\log|S|/\epsilon^2\right)}$, with the property that

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|f(\mathbf{x})-f(\mathbf{y})\|_2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \tag{1.1}$$

holds $\forall \mathbf{x} \in S$ and $\forall \mathbf{y} \in \mathbb{R}^N$.

In this paper we generalize Narayanan and Nelson's theorem for finite sets to also hold for infinite subsets $S \subset \mathbb{R}^N$, and then give a specialized variant for the case where the infinite subset $S \subset \mathbb{R}^N$ in question is a compact and smooth submanifold of $\mathbb{R}^N$. As we shall see below, generalizing this result requires us to both alter the bi-Lipschitz extension methods of [18] as well as to replace the use of embedding techniques utilizing cardinality in [19] with different JL-type embedding methods involving alternate measures of set complexity which remain meaningful for infinite sets (i.e., the Gaussian width of the unit secants of the set

*Department of Mathematics, and Department of Computational Mathematics, Science and Engineering (CMSE), Michigan State University, East Lansing, MI 48824. **E-mail:** iwenmark@msu.edu

†Department of Mathematics, Michigan State University, East Lansing, MI 48824. **E-mail:** roachma3@msu.edu

S in question). In the special case where S is a submanifold of $\mathbb{R}^N$, recent results bounding the Gaussian widths of the unit secants of such sets in terms of other fundamental geometric quantities (e.g., their reach, dimension, volume, etc.) [13] can then be brought to bear in order to produce terminal manifold embeddings of S into $\mathbb{R}^m$ satisfying (1.1) with m near-optimally small.

Note that a non-trivial terminal embedding, f, of S satisfying (1.1) for all $\mathbf{x} \in S$ and $\mathbf{y} \in \mathbb{R}^N$ must be nonlinear. In contrast, prior work on bi-Lipschitz maps of submanifolds of $\mathbb{R}^N$ into lower dimensional Euclidean space in the mathematical data science literature have all utilized *linear* maps (see, e.g., [1, 7, 13]). As a result, it is impossible for such previously considered linear maps to serve as terminal embeddings of submanifolds of $\mathbb{R}^N$ into lower-dimensional Euclidean space without substantial modification. Another way of viewing the work carried out herein is that it constructs outer bi-Lipschitz extensions of such prior linear JL embeddings of manifolds in a way that effectively preserves their near-optimal embedding dimension in the final resulting extension. Motivating applications of terminal embeddings of submanifolds of $\mathbb{R}^N$ related to compressive classification via manifold models [5] are discussed next.

1.1 Universally Accurate Compressive Classification via Noisy Manifold Data

It is one of the sad facts of life that most everyone eventually comes to accept: everything living must eventually die, you can't always win, you aren't always right, and – worst of all to the most dedicated of data scientists – there is always noise contaminating your datasets. Nevertheless, there are mitigating circumstances and achievable victories implicit in every statement above – most pertinently here, there are mountains of empirical evidence that noisy training data still permits accurate learning. In particular, when the noise level is not too large, the mere existence of a low-dimensional data model which only approximately fits your noisy training data can still allow for successful, e.g., nearest-neighbor classification using only a highly compressed version of your original training dataset (even when you know very little about the model specifics) [5]. Better quantifying these empirical observations in the context of low-dimensional manifold models is the primary motivation for our main result below.

For example, let $\mathcal{M} \subset \mathbb{R}^N$ be a d-dimensional submanifold of $\mathbb{R}^N$ (our data model), fix $\delta \in \mathbb{R}^+$ (our effective noise level), and choose $T \subseteq \text{tube}(\delta, \mathcal{M}) := \{\mathbf{x} \mid \exists \mathbf{y} \in \mathcal{M} \text{ with } \|\mathbf{x} - \mathbf{y}\|_2 \leq \delta\}$ (our "noisy" and potentially high-dimensional training data). Fix $\epsilon \in (0, 1)$. For a terminal embedding $f : \mathbb{R}^N \to \mathbb{R}^m$ of $\mathcal{M}$ as per (1.1), one can see that

$$(1-\epsilon)\|\mathbf{z} - \mathbf{t}\|_2 - 2(1-\epsilon)\delta \leq \|f(\mathbf{z}) - f(\mathbf{t})\|_2 \leq (1+\epsilon)\|\mathbf{z} - \mathbf{t}\|_2 + 2(1+\epsilon)\delta \tag{1.2}$$

will hold simultaneously for all $\mathbf{z} \in \mathbb{R}^N$ and $\mathbf{t} \in T$, *where f has an embedding dimension that only depends on the geometric properties of $\mathcal{M}$ (and not necessarily on $|T|$)*.[1] Thus, if T includes a sufficiently dense external cover of $\mathcal{M}$, then f will allow us to approximate the distance of all $\mathbf{z} \in \mathbb{R}^N$ to $\mathcal{M}$ in the compressed embedding space via the estimator

$$\tilde{d}(f(\mathbf{z}), f(T)) := \inf_{\mathbf{t} \in T} \|f(\mathbf{z}) - f(\mathbf{t})\|_2 \approx d(\mathbf{z}, \mathcal{M}) := \inf_{\mathbf{y} \in \mathcal{M}} \|\mathbf{z} - \mathbf{y}\|_2 \tag{1.3}$$

up to $\mathcal{O}(\delta)$-error. As a result, if one has noisy data from two disjoint manifolds $\mathcal{M}_1, \mathcal{M}_2 \subset \mathbb{R}^N$, one can use this compressed $\tilde{d}$ estimator to correctly classify all data $\mathbf{z} \in \text{tube}(\delta, \mathcal{M}_1) \bigcup \text{tube}(\delta, \mathcal{M}_2)$ as being in either $T_1 := \text{tube}(\delta, \mathcal{M}_1)$ (class 1) or $T_2 := \text{tube}(\delta, \mathcal{M}_2)$ (class 2) as long as $\inf_{\mathbf{x} \in T_1, \mathbf{y} \in T_2} \|\mathbf{x} - \mathbf{y}\|_2$ is sufficiently large. In short, terminal manifold embeddings demonstrate that accurate compressive nearest-neighbor classification based on noisy manifold training data is always possible as long as the manifolds in question are sufficiently far apart (though not necessarily separable from one another by, e.g., a hyperplane, etc.).

Note that in the discussion above we may in fact take $T = \text{tube}(\delta, \mathcal{M})$. In that case (1.2) will hold simultaneously for all $\mathbf{z} \in \mathbb{R}^N$ and $(\mathbf{t}, \delta) \in \mathbb{R}^N \times \mathbb{R}^+$ with $\mathbf{t} \in \text{tube}(\delta, \mathcal{M})$ so that $f : \mathbb{R}^N \to \mathbb{R}^m$ will approximately preserve the distances of all points $\mathbf{z} \in \mathbb{R}^N$ to $\text{tube}(\delta, \mathcal{M})$ up to errors on the order of

[1] One can prove (1.2) by comparing both $\mathbf{z}$ and $\mathbf{t}$ to a point $\mathbf{x}_\mathbf{t} \in \mathcal{M}$ satisfying $\|\mathbf{t} - \mathbf{x}_\mathbf{t}\|_2 \leq \delta$ via several applications of the (reverse) triangle inequality.

$\mathcal{O}(\epsilon)d(\mathbf{z}, \text{tube}(\delta, \mathcal{M})) + \mathcal{O}(\delta)$ for all $\delta \in \mathbb{R}^+$. This is in fact rather remarkable when one recalls that the best achievable embedding dimension, m, here only depends on the geometric properties of the low-dimensional manifold $\mathcal{M}$ (see Theorem 1.1 for a detailed accounting of these dependences).

We further note that alternate applications of Theorem 3.2 (on which Theorem 1.1 depends) involving other data models are also possible. As a more explicit second example, suppose that $\mathcal{M}$ is a union of n d-dimensional affine subspaces so that its unit secants, $S_{\mathcal{M}}$ defined as per (2.1), are contained in the union of at most $\binom{n}{2} + n$ unit spheres $\subset \mathbb{S}^{N-1}$, each of dimension at most $2d+1$. The Gaussian width (see Definition 2.1) of $S_{\mathcal{M}}$ can then be upper-bounded by $C\sqrt{d + \log n}$ using standard techniques, where $C \in \mathbb{R}^+$ is an absolute constant. An application of Theorem 3.2 now guarantees the existence of a terminal embedding $f : \mathbb{R}^N \to \mathbb{R}^{\mathcal{O}\left(\frac{d+\log n}{\epsilon^2}\right)}$ which will allow approximate nearest subspace queries to be answered for any input point $\mathbf{z} \in \mathbb{R}^N$ using only $f(\mathbf{z})$ in the compressed $\mathcal{O}\left(\frac{d+\log n}{\epsilon^2}\right)$-dimensional space. Even more specifically, if we choose, e.g., $\mathcal{M}$ to consist of all at most s-sparse vectors in $\mathbb{R}^N$ (i.e., so that $\mathcal{M}$ is the union of $n = \binom{N}{s}$ subspaces of $\mathbb{R}^N$), we can now see that Theorem 3.2 guarantees the existence of a deterministic compressed estimator (1.3) which allows for the accurate approximation of the best s-term approximation error $\inf_{\mathbf{y} \in \mathbb{R}^N \text{ at most } s \text{ sparse}} \|\mathbf{z} - \mathbf{y}\|_2$ for all $\mathbf{z} \in \mathbb{R}^N$ using only $f(\mathbf{z}) \in \mathbb{R}^{\mathcal{O}(s \log(N/s))}$ as input. Note that this is only possible due to the non-linearity of f herein. In, e.g., the setting of classical compressive sensing theory where f must be linear it is known that such good performance is impossible [4, Section 5].

1.2 The Main Result and a Brief Outline of Its Proof

The following theorem is proven in Section 4. Given a low-dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^N$ it establishes the existence of a function $f : \mathbb{R}^N \to \mathbb{R}^m$ with $m \ll N$ that approximately preserves the Euclidean distances from all points in $\mathbb{R}^N$ to all points in $\mathcal{M}$. As a result, it guarantees the existence of a low-dimensional embedding which will, e.g., always allow for the correct compressed nearest-neighbor classification of images living near different well separated submanifolds of Euclidean space.

Theorem 1.1 (The Main Result)**.** *Let $\mathcal{M} \hookrightarrow \mathbb{R}^N$ be a compact d-dimensional submanifold of $\mathbb{R}^N$ with boundary $\partial\mathcal{M}$, finite reach $\tau_{\mathcal{M}}$ (see Definition 2.2), and volume $V_{\mathcal{M}}$. Enumerate the connected components of $\partial\mathcal{M}$ and let τ_i be the reach of the i^{th} connected component of $\partial\mathcal{M}$ as a submanifold of $\mathbb{R}^N$. Set $\tau := \min_i\{\tau_{\mathcal{M}}, \tau_i\}$, let $V_{\partial\mathcal{M}}$ be the volume of $\partial\mathcal{M}$, and denote the volume of the d-dimensional Euclidean ball of radius 1 by ω_d. Next,*

1. *if $d = 1$, define $\alpha_{\mathcal{M}} := \frac{20V_{\mathcal{M}}}{\tau} + V_{\partial\mathcal{M}}$, else*
2. *if $d \geq 2$, define $\alpha_{\mathcal{M}} := \frac{V_{\mathcal{M}}}{\omega_d}\left(\frac{41}{\tau}\right)^d + \frac{V_{\partial\mathcal{M}}}{\omega_{d-1}}\left(\frac{81}{\tau}\right)^{d-1}$.*

Finally, fix $\epsilon \in (0,1)$ and define

$$\beta_{\mathcal{M}} := \left(\alpha_{\mathcal{M}}^2 + 3^d\alpha_{\mathcal{M}}\right). \tag{1.4}$$

Then, there exists a map $f : \mathbb{R}^N \to \mathbb{C}^m$ with $m \leq c\left(\ln\left(\beta_{\mathcal{M}}\right) + 4d\right)/\epsilon^2$ that satisfies

$$\left|\|f(\mathbf{x}) - f(\mathbf{y})\|_2^2 - \|\mathbf{x} - \mathbf{y}\|_2^2\right| \leq \epsilon\|\mathbf{x} - \mathbf{y}\|_2^2 \tag{1.5}$$

for all $\mathbf{x} \in \mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$. Here $c \in \mathbb{R}^+$ is an absolute constant independent of all other quantities.

Proof. See Section 4. □

The remainder of the paper is organized as follows. In Section 2 we review notation and state a result from [13] that bounds the Gaussian width of the unit secants of a given submanifold of $\mathbb{R}^N$ in terms of geometric quantities of the original submanifold. Next, in Section 3 we prove an optimal terminal embedding result for arbitrary subsets of $\mathbb{R}^N$ in terms of the Gaussian widths of their unit secants by generalizing results from

the computer science literature concerning finite sets [18, 19]. See Theorem 3.2 therein. We then combine results from Sections 2 and 3 in order to prove our main theorem in Section 4. Finally, in Section 5 we conclude by demonstrating that terminal embeddings allow for more accurate compressive nearest neighbor classification than standard linear embeddings in practice.

2 Notation and Preliminaries

Below $B_{\ell^2}^N(\mathbf{x}, \gamma)$ will denote the open Euclidean ball around $\mathbf{x}$ of radius γ in $\mathbb{R}^N$. Given an arbitrary subset $S \subset \mathbb{R}^N$, we will further define $-S := \{-\mathbf{x} \mid \mathbf{x} \in S\}$ and $S \pm S := \{\mathbf{x} \pm \mathbf{y} \mid \mathbf{x}, \mathbf{y} \in S\}$. Finally, for a given $T \subset \mathbb{R}^N$ we will also let $\overline{T}$ denote its closure, and further define the normalization operator $U : \mathbb{R}^N \setminus \{\mathbf{0}\} \to \mathbb{S}^{N-1}$ to be such that $U(\mathbf{x}) := \mathbf{x}/\|\mathbf{x}\|_2$. With this notation in hand we can then define the **unit secants of** $T \subset \mathbb{R}^N$ to be

$$S_T := \overline{U\left((T-T) \setminus \{\mathbf{0}\}\right)} = \overline{\left\{ \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2} \;\middle|\; \mathbf{x}, \mathbf{y} \in T,\ \mathbf{x} \neq \mathbf{y} \right\}}. \tag{2.1}$$

Note that S_T is always a compact subset of the unit sphere $\mathbb{S}^{N-1} \subset \mathbb{R}^N$, and that $S_T = -S_T$.

Herein we will call a matrix $A \in \mathbb{C}^{m \times N}$ an *ϵ-JL **map** of a set $T \subset \mathbb{R}^N$ into $\mathbb{C}^m$* if

$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2$$

holds for all $\mathbf{x} \in T$. Note that this is equivalent to $A \in \mathbb{C}^{m \times N}$ having the property that

$$\sup_{\mathbf{x} \in T \setminus \{\mathbf{0}\}} \left| \|A(\mathbf{x}/\|\mathbf{x}\|_2)\|_2^2 - 1 \right| = \sup_{\mathbf{x} \in U(T)} \left| \|A\mathbf{x}\|_2^2 - 1 \right| \leq \epsilon,$$

where $U(T) \subset \mathbb{R}^N$ is the normalized version of $T \setminus \{\mathbf{0}\} \subset \mathbb{R}^{\mathbf{N}}$ defined as above. Furthermore, we will say that a matrix $A \in \mathbb{C}^{m \times n}$ is an *ϵ-JL **embedding** of a set $T \subset \mathbb{R}^n$ into $\mathbb{C}^m$* if A is an ϵ-JL map of

$$T - T := \{\mathbf{x} - \mathbf{y} \mid \mathbf{x}, \mathbf{y} \in T\}$$

into $\mathbb{C}^m$. Here we will be working with random matrices which will embed any fixed set T of bounded size (measured with respect to, e.g., Gaussian Width [23]) with high probability. Such matrix distributions are often called **oblivious** and discussed as randomized embeddings in the absence of any specific set T since their embedding quality can be determined independently of any properties of a given set T beyond its size. In particular, the class of oblivious **sub-Gaussian random matrices** having independent, isotropic, and sub-Gaussian rows will receive special attention below.

2.1 Some Common Measures of Set Size and Complexity with Associated Bounds

We will denote the cardinality of a finite set T by $|T|$. For a (potentially infinite) set $T \subset \mathbb{R}^N$ we define its **radius** and **diameter** to be

$$\operatorname{rad}(T) := \sup_{\mathbf{x} \in T} \|\mathbf{x}\|_2$$

and

$$\operatorname{diam}(T) := \operatorname{rad}(T - T) = \sup_{\mathbf{x}, \mathbf{y} \in T} \|\mathbf{x} - \mathbf{y}\|_2,$$

respectively. Given a value $\delta \in \mathbb{R}^+$, a **δ-cover of** T (also sometimes called a **δ-net of** T) will be a subset $S \subset T$ such that the following holds

$$\forall \mathbf{x} \in T\ \exists \mathbf{y} \in S \text{ such that } \|\mathbf{x} - \mathbf{y}\|_2 \leq \delta.$$

The **δ-covering number of** T, denoted by $\mathcal{N}(T, \delta) \in \mathbb{N}$, is then the smallest achievable cardinality of a δ-cover of T. Finally, the **Gaussian width** of a set T is defined as follows.

Definition 2.1. *(**Gaussian Width** [23, Definition 7.5.1]). The Gaussian width of a set $T \subset \mathbb{R}^N$ is*

$$w(T) := \mathbb{E} \sup_{\mathbf{x} \in T} \langle \mathbf{g}, \mathbf{x} \rangle$$

where $\mathbf{g}$ is a random vector with N independent and identically distributed (i.i.d.) mean 0 and variance 1 Gaussian entries. For a list of useful properties of the Gaussian width we refer the reader to [23, Proposition 7.5.2].

Finally, reach is an extrinsic parameter of a subset S of Euclidean space defined based on how far away points can be from S while still having a unique closest point in S [8, 22]. The following formal definition of reach utilizes the Euclidean distance d between a given point $\mathbf{x} \in \mathbb{R}^N$ and subset $S \subset \mathbb{R}^N$.

Definition 2.2. *(**Reach** [8, Definition 4.1]). For a subset $S \subset \mathbb{R}^N$ of Euclidean space, the reach τ_S is*

$$\tau_S := \sup \left\{ t \geq 0 \;\middle|\; \forall \mathbf{x} \in \mathbb{R}^n \text{ such that } d(\mathbf{x}, S) < t, \mathbf{x} \text{ has a unique closest point in } S \right\}.$$

The following theorem is a restatement of Theorem 20 in [13]. It bounds the Gaussian width of a smooth submanifold of $\mathbb{R}^N$ in terms of its dimension, reach, and volume.

Theorem 2.1 (Gaussian Width of the Unit Secants of a Submanifold of $\mathbb{R}^N$, Potentially with Boundary)**.** *Let $\mathcal{M} \hookrightarrow \mathbb{R}^N$ be a compact d-dimensional submanifold of $\mathbb{R}^N$ with boundary $\partial\mathcal{M}$, finite reach $\tau_{\mathcal{M}}$, and volume $V_{\mathcal{M}}$. Enumerate the connected components of $\partial\mathcal{M}$ and let τ_i be the reach of the i^{th} connected component of $\partial\mathcal{M}$ as a submanifold of $\mathbb{R}^N$. Set $\tau := \min_i\{\tau_{\mathcal{M}}, \tau_i\}$, let $V_{\partial\mathcal{M}}$ be the volume of $\partial\mathcal{M}$, and denote the volume of the d-dimensional Euclidean ball of radius 1 by ω_d. Next,*

1. *if $d = 1$, define $\alpha_{\mathcal{M}} := \frac{20V_{\mathcal{M}}}{\tau} + V_{\partial\mathcal{M}}$, else*
2. *if $d \geq 2$, define $\alpha_{\mathcal{M}} := \frac{V_{\mathcal{M}}}{\omega_d}\left(\frac{41}{\tau}\right)^d + \frac{V_{\partial\mathcal{M}}}{\omega_{d-1}}\left(\frac{81}{\tau}\right)^{d-1}$.*

Finally, define

$$\beta_{\mathcal{M}} := \left(\alpha_{\mathcal{M}}^2 + 3^d \alpha_{\mathcal{M}}\right). \tag{2.2}$$

Then, the Gaussian width of $\overline{U\left((\mathcal{M} - \mathcal{M}) \setminus \{\mathbf{0}\}\right)}$ satisfies

$$w\left(S_{\mathcal{M}}\right) = w\left(\overline{U\left((\mathcal{M} - \mathcal{M}) \setminus \{\mathbf{0}\}\right)}\right) \leq 8\sqrt{2}\sqrt{\ln\left(\beta_{\mathcal{M}}\right) + 4d}.$$

With this Gaussian width bound in hand we can now begin the proof of our main result. The approach will be to combine Theorem 2.1 above with general theorems concerning the existence of outer bi-Lipschitz extensions of ϵ-JL embeddings of arbitrary subsets of $\mathbb{R}^N$ into lower-dimensional Euclidean space. These general existence theorems are proven in the next section.

3 The Main Bi-Lipschitz Extension Results and Their Proofs

Our first main technical result guarantees that any given JL map Φ of a special subset of $\mathbb{S}^{N-1}$ related to $\mathcal{M}$ will not only be a bi-Lipschitz map from $\mathcal{M} \subset \mathbb{R}^N$ into a lower dimensional Euclidean space $\mathbb{R}^m$, but will also have an outer bi-Lipschitz extension into $\mathbb{R}^{m+1}$. It is useful as a means of extending particular (structured) JL maps Φ of special interest in the context of, e.g., saving on memory costs [12].

Theorem 3.1. *Let $\mathcal{M} \subset \mathbb{R}^N$, $\epsilon \in (0, 1)$, and suppose that $\Phi \in \mathbb{C}^{m \times N}$ is an $\left(\frac{\epsilon^2}{2304}\right)$-JL map of $S_{\mathcal{M}} + S_{\mathcal{M}}$ into $\mathbb{C}^m$. Then, there exists an outer bi-Lipschitz extension of $\Phi : \mathcal{M} \to \mathbb{C}^m$, $f : \mathbb{R}^N \to \mathbb{C}^{m+1}$, with the property that*

$$\left| \|f(\mathbf{x}) - f(\mathbf{y})\|_2^2 - \|\mathbf{x} - \mathbf{y}\|_2^2 \right| \leq \epsilon \|\mathbf{x} - \mathbf{y}\|_2^2$$

holds for all $\mathbf{x} \in \mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$.

Proof. See Section 3.4. □

Looking at Theorem 3.1 we can see that an $\left(\frac{\epsilon^2}{2304}\right)$-JL map of $S_{\mathcal{M}} + S_{\mathcal{M}}$ is required in order to achieve the outer extension f of interest. This result is sub-optimal in two respects. First, the constant factor $1/2304$ is certainly not tight and can likely be improved substantially. More importantly though is the fact that ϵ is squared in the required map distortion which means that the terminal embedding dimension, $m+1$, will have to scale sub-optimally in ϵ (see Remark 3.1 below for details). Unfortunately, this is impossible to rectify when extending arbitrary maps Φ (see, e.g., [18]). For sub-gaussian Φ an improvement is in fact possible, however, which is the subject of our second main technical result just below. Using specialized theory for sub-gaussian matrices it demonstrates the existence of terminal JL embeddings for arbitrary subsets of $\mathbb{R}^N$ which achieve an optimal terminal embedding dimension up to constants.

Theorem 3.2. *Let $\mathcal{M} \subset \mathbb{R}^N$ and $\epsilon \in (0,1)$. There exists a map $f : \mathbb{R}^N \to \mathbb{C}^m$ with $m \leq c\left(\frac{w(S_{\mathcal{M}})}{\epsilon}\right)^2$ that satisfies*

$$\left| \|f(\mathbf{x}) - f(\mathbf{y})\|_2^2 - \|\mathbf{x} - \mathbf{y}\|_2^2 \right| \leq \epsilon \|\mathbf{x} - \mathbf{y}\|_2^2 \tag{3.1}$$

for all $\mathbf{x} \in \mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$. Here $c \in \mathbb{R}^+$ is an absolute constant independent of all other quantities.

Proof. See Section 3.5. □

To see the optimality of the terminal embedding dimension m provided by Theorem 3.2 we note that functions f which satisfy (3.1) for all $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ must in fact generally scale quadratically in both $w(S_{\mathcal{M}})$ and $1/\epsilon$ (see [11, Theorem 7] and [16]). We will now begin proving supporting results for both of the main technical theorems above. The first supporting results pertain to the so-called **convex hull distortion** of a given linear ϵ-JL map.

3.1 All Linear ϵ-JL Maps Provide $\mathcal{O}(\sqrt{\epsilon})$-Convex Hull Distortion

A crucial component involved in proving our main results involves the approximate norm preservation of all points in the convex hull of a given set bounded set $S \subset \mathbb{R}^N$. Recall that the **convex hull of** $S \subset \mathbb{C}^N$ is

$$\text{conv}(S) := \bigcup_{j=1}^{\infty} \left\{ \sum_{\ell=1}^{j} \alpha_\ell \mathbf{x}_\ell \;\middle|\; \mathbf{x}_1, \ldots, \mathbf{x}_j \in S,\, \alpha_1, \ldots, \alpha_j \in [0,1] \text{ s.t. } \sum_{\ell=1}^{j} \alpha_\ell = 1 \right\}.$$

The next theorem states that each point in the convex hull of $S \subset \mathbb{R}^N$ can be expressed as a convex combination of at most $N+1$ points from S. Hence, the convex hulls of subsets of $\mathbb{R}^N$ are actually a bit simpler than they first appear.

Theorem 3.3 (Carathéadory, see, e.g., [2])**.** *Given $S \in \mathbb{R}^N$, $\forall \mathbf{x} \in$ conv(S), $\exists \mathbf{y}_1, \ldots, \mathbf{y}_{\tilde{N}}$, $\tilde{N} = \min(|S|, N+1)$, such that $\mathbf{x} = \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{y}_\ell$ for some $\alpha_1, \ldots, \alpha_{\tilde{N}} \in [0,1]$, $\sum_{\ell=1}^{\tilde{N}} \alpha_\ell = 1$.*

Finally, we say that a matrix $\Phi \in \mathbb{C}^{m \times N}$ provides **ϵ-convex hull distortion for** $S \subset \mathbb{R}^N$ if

$$|\|\Phi \mathbf{x}\|_2 - \|\mathbf{x}\|_2| \leq \epsilon$$

holds for all $\mathbf{x} \in \text{conv}(S)$. The main result of this subsection states that all linear ϵ-JL maps can provide ϵ-convex hull distortion for the unit secants of any given set. In particular, we have the following theorem which generalizes arguments in [18] for finite sets to arbitrary and potentially infinite sets.

Theorem 3.4. *Let $\mathcal{M} \subset \mathbb{R}^N$, $\epsilon \in (0,1)$, and suppose that $\Phi \in \mathbb{C}^{m \times N}$ is an $\left(\frac{\epsilon^2}{4}\right)$-JL map of $S_{\mathcal{M}} + S_{\mathcal{M}}$ into $\mathbb{C}^m$. Then, Φ will also provide ϵ-convex hull distortion for $S_{\mathcal{M}}$.*

The proof of Theorem 3.4 depends on two intermediate lemmas. The first lemma is a slight modification of Lemma 3 in [12].

Lemma 3.1. *Let $S \subset \mathbb{R}^N$ and $\epsilon \in (0,1)$. Then, an ϵ-JL map $\Phi \in \mathbb{C}^{m\times N}$ of the set*

$$S' = \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \mid \mathbf{x}, \mathbf{y} \in S \right\}$$

will satisfy

$$|\Re\left(\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle\right) - \langle \mathbf{x}, \mathbf{y}\rangle| \leq 2\epsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

$\forall \mathbf{x}, \mathbf{y} \in S$.

Proof. If $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ the inequality holds trivially. Thus, suppose $\mathbf{x}, \mathbf{y} \neq 0$. Consider the normalizations $\mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \mathbf{v} = \frac{\mathbf{y}}{\|\mathbf{y}\|_2}$. The polarization identities for complex/real inner products imply that

$$\begin{aligned}
|\Re\left(\langle \Phi\mathbf{u}, \Phi\mathbf{v}\rangle\right) - \langle \mathbf{u}, \mathbf{v}\rangle| &= \frac{1}{4}\left|\Re\left(\sum_{\ell=0}^{3} i^\ell \left\|\Phi\mathbf{u} + i^\ell \Phi\mathbf{v}\right\|_2^2\right) - \left(\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2\right)\right| \\
&= \frac{1}{4}\left|\left(\|\Phi\mathbf{u}+\Phi\mathbf{v}\|_2^2 - \|\Phi\mathbf{u}-\Phi\mathbf{v}\|_2^2\right) - \left(\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2\right)\right| \\
&\leq \frac{1}{4}\left(\left|\|\Phi\mathbf{u}+\Phi\mathbf{v}\|_2^2 - \|\mathbf{u}+\mathbf{v}\|_2^2\right| + \left|\|\Phi\mathbf{u}-\Phi\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2\right|\right) \\
&\leq \frac{\epsilon}{4}\left(\|\mathbf{u}+\mathbf{v}\|_2^2 + \|\mathbf{u}-\mathbf{v}\|_2^2\right) \leq \frac{\epsilon}{2}\left(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2\right)^2 \leq 2\epsilon.
\end{aligned}$$

The result now follows by multiplying the inequality through by $\|\mathbf{x}\|_2\|\mathbf{y}\|_2$. □

Next, we see that and linear ϵ-JL maps are capable of preserving the angles between the elements of the convex hull of any bounded subset $S \subset \mathbb{R}^N$.

Lemma 3.2. *Suppose $S \subset \overline{B_{\ell^2}^N(\mathbf{0}, \gamma)}$ and $\epsilon \in (0,1)$. Let $\Phi \in \mathbb{C}^{m\times N}$ be an $\left(\frac{\epsilon}{2\gamma^2}\right)$-JL map of the set S' defined as in Lemma 3.1 into $\mathbb{C}^m$. Then*

$$|\Re\left(\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle\right) - \langle \mathbf{x}, \mathbf{y}\rangle| \leq \epsilon$$

holds $\forall \mathbf{x}, \mathbf{y} \in$ conv(S).

Proof. Let $\mathbf{x}, \mathbf{y} \in \text{conv}(S)$. By Theorem 3.3, $\exists \{\mathbf{y}_i\}_{i=1}^{\tilde{N}}, \{\mathbf{x}_i\}_{i=1}^{\tilde{N}} \subset S$ and $\{\alpha_\ell\}_{\ell=1}^{\tilde{N}}, \{\beta_\ell\}_{\ell=1}^{\tilde{N}} \subset [0,1]$ with $\sum_{\ell=1}^{\tilde{N}} \alpha_\ell = \sum_{\ell=1}^{\tilde{N}} \beta_\ell = 1$ such that

$$\mathbf{x} = \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{x}_\ell, \text{ and } \mathbf{y} = \sum_{\ell=1}^{\tilde{N}} \beta_\ell \mathbf{y}_\ell.$$

Hence, by Lemma 3.1 we have that

$$\begin{aligned}
|\Re\left(\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle\right) - \langle \mathbf{x}, \mathbf{y}\rangle| &= \left|\sum_{\ell=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}} \alpha_\ell\beta_j \left(\Re\left(\langle \Phi\mathbf{x}_\ell, \Phi\mathbf{y}_j\rangle\right) - \langle \mathbf{x}_\ell, \mathbf{y}_j\rangle\right)\right| \\
&\leq 2\sum_{\ell=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}} \alpha_\ell\beta_j \left(\frac{\epsilon}{2\gamma^2}\right) \|\mathbf{x}_\ell\|_2 \|\mathbf{y}_j\|_2 \\
&\leq \epsilon \left(\sum_{\ell=1}^{\tilde{N}} \alpha_\ell\right)\left(\sum_{j=1}^{\tilde{N}} \beta_j\right) \;=\; \epsilon.
\end{aligned}$$

Here we have also used the mapping error $\left(\frac{\epsilon}{2\gamma^2}\right)$ and the fact that all norms of vectors in this case will be less than γ. □

We are now prepared to prove Theorem 3.4.

3.1.1 Proof of Theorem 3.4

Applying Lemma 3.2 with $S = S_{\mathcal{M}} = S_{\mathcal{M}} \cup -S_{\mathcal{M}}$, we note that $S' = S_{\mathcal{M}} + S_{\mathcal{M}} = (S_{\mathcal{M}} \cup -S_{\mathcal{M}}) + (S_{\mathcal{M}} \cup -S_{\mathcal{M}})$ since $S \subset \mathbb{S}^{N-1}$. Furthermore, $\gamma = 1$ in this case. Hence, $\Phi \in \mathbb{R}^{m \times N}$ being an $\left(\frac{\epsilon^2}{4}\right)$-JL map of $S_{\mathcal{M}} + S_{\mathcal{M}}$ into $\mathbb{R}^m$ implies that

$$|\Re\left(\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle\right) - \langle \mathbf{x}, \mathbf{y}\rangle| \leq \frac{\epsilon^2}{2} \tag{3.2}$$

holds $\forall \mathbf{x}, \mathbf{y} \in \operatorname{conv}(S_{\mathcal{M}}) \subset \overline{B^N_{\ell^2}(\mathbf{0}, 1)}$. In particular, (3.2) with $\mathbf{x} = \mathbf{y}$ implies that

$$|\|\Phi\mathbf{x}\|_2 - \|\mathbf{x}\|_2|\,|\|\Phi\mathbf{x}\|_2 + \|\mathbf{x}\|_2| \;=\; \left|\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \epsilon^2/2.$$

Noting that $|\|\Phi\mathbf{x}\|_2 + \|\mathbf{x}\|_2| \geq \|\mathbf{x}\|_2$ we can see that the desired result holds automatically if $\|\mathbf{x}\|_2 \geq \epsilon/2$. Thus, it suffices to assume that that $\|\mathbf{x}\|_2 < \epsilon/2$, but then we are also finished since $|\|\Phi\mathbf{x}\|_2 - \|\mathbf{x}\|_2| \leq \max\{\|\mathbf{x}\|_2, \|\Phi\mathbf{x}\|_2\} \leq \sqrt{\|\mathbf{x}\|_2^2 + \epsilon^2/2} < \frac{\sqrt{3}}{2}\epsilon$ will hold in that case.

Remark 3.1. *Though Theorem 3.4 holds for arbitrary linear maps, we note that it has suboptimal dependence on the distortion parameter ϵ. In particular, a linear $\left(\frac{\epsilon^2}{4}\right)$-JL map of an arbitrary set will generally embed that set into $\mathbb{C}^m$ with $m = \Omega(1/\epsilon^4)$ [16]. However, it has been shown in [19] that sub-Gaussian matrices will behave better with high probability, allowing for outer bi-Lipschitz extensions of JL-embeddings of finite sets into $\mathbb{R}^m$ with $m = \mathcal{O}(1/\epsilon^2)$. In the next subsection we generalize those better scaling results for sub-Gaussian random matrices to (potentially) infinite sets.*

3.2 Sub-Gaussian Matrices and ϵ-Convex Hull Distortion for Infinite Sets

Motivated by results in [19] for finite sets which achieve optimal dependence on the distortion parameter ϵ for sub-Gaussian matrices, in this section we will do the same for infinite sets using results from [23]. Our main tool will be the following result (see also [13, Theorem 4]).

Theorem 3.5 (See Theorem 9.1.1 and Exercise 9.1.8 in [23])**.** *Let Φ be $m \times N$ matrix whose rows are independent, isotropic, and sub-Gaussian random vectors in $\mathbb{R}^N$. Let $p \in (0,1)$ and $S \subset \mathbb{R}^N$. Then there exists a constant c depending only on the distribution of the rows of Φ such that*

$$\sup_{\mathbf{x} \in S} \left|\|\Phi\mathbf{x}\|_2 - \sqrt{m}\|\mathbf{x}\|_2\right| \leq c\left[w(S) + \sqrt{\ln(2/p)} \cdot \operatorname{rad}(S)\right]$$

holds with probability at least $1 - p$.

The main result of this section is a simple consequence of Theorem 3.5 together with standard results concerning Gaussian widths [23, Proposition 7.5.2].

Corollary 3.1. *Let $\mathcal{M} \subset \mathbb{R}^N$, $\epsilon, p \in (0,1)$, and $\Phi \in \mathbb{R}^{m \times N}$ be an $m \times N$ matrix whose rows are independent, isotropic, and sub-Gaussian random vectors in $\mathbb{R}^N$. Furthermore, suppose that*

$$m \geq \frac{c'}{\epsilon^2}\left(w\left(S_{\mathcal{M}}\right) + \sqrt{\ln(2/p)}\right)^2,$$

where c' is a constant depending only on the distribution of the rows of Φ. Then, with probability at least $1 - p$ the random matrix $\frac{1}{\sqrt{m}}\Phi$ will simultaneously be both an ϵ-JL embedding of $\mathcal{M}$ into $\mathbb{R}^m$ and also provide ϵ-convex hull distortion for $S_{\mathcal{M}}$.

Proof. We apply Theorem 3.5 to $S = \operatorname{conv}\left(S_{\mathcal{M}}\right)$. In doing so we note that $w\left(\operatorname{conv}\left(S_{\mathcal{M}}\right)\right) = w\left(S_{\mathcal{M}}\right)$ [23, Proposition 7.5.2], and that $\operatorname{rad}\left(\operatorname{conv}\left(S_{\mathcal{M}}\right)\right) = 1$ since $\operatorname{conv}\left(S_{\mathcal{M}}\right) \subseteq \overline{B^N_{\ell^2}(\mathbf{0}, 1)}$. The result will be that $\frac{1}{\sqrt{m}}\Phi$ provides ϵ-convex hull distortion for $S_{\mathcal{M}}$ as long as $c' \geq c^2$. Next, we note that providing ϵ-convex hull distortion for $S_{\mathcal{M}}$ implies that $\frac{1}{\sqrt{m}}\Phi$ will also approximately preserve the ℓ_2-norms of all the unit vectors

in $S_{\mathcal{M}} \subset \operatorname{conv}(S_{\mathcal{M}})$. In particular, $\frac{1}{\sqrt{m}}\Phi$ will be a 3ϵ-JL map of $S_{\mathcal{M}}$ into $\mathbb{R}^m$, which in turn implies that $\frac{1}{\sqrt{m}}\Phi$ will also be a 3ϵ-JL embedding of $\mathcal{M}-\mathcal{M}$ into $\mathbb{R}^m$ by linearity/rescaling. Adjusting the constant c' to account for the additional factor of 3 now yields the stated result. □

We are now prepared to prove our general theorems regarding outer bi-Lipschitz extensions of JL-embeddings of potentially infinite sets.

3.3 Outer Bi-Lipschitz Extension Results for JL-embeddings of General Sets

Before we can prove our final results for general sets we will need two supporting lemmas. They are adapted from the proofs of analogous results in [18, 19] for finite sets.

Lemma 3.3. *Let $\mathcal{M} \subset \mathbb{R}^N$, $\epsilon \in (0,1)$, and suppose that $\Phi \in \mathbb{C}^{m\times N}$ provides ϵ-convex hull distortion for $S_{\mathcal{M}}$. Then, there exists a function $g : \mathbb{R}^N \to \mathbb{C}^m$ such that*

$$|\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}\rangle\right) - \langle \mathbf{y}, \mathbf{x}\rangle| \leq 2\epsilon\|\mathbf{y}\|_2\|\mathbf{x}\|_2 \tag{3.3}$$

holds for all $\mathbf{x} \in \overline{\mathcal{M}} - \overline{\mathcal{M}}$ and $\mathbf{y} \in \mathbb{R}^N$.

Proof. First, we note that (3.3) holds trivially for $\mathbf{y} = \mathbf{0}$ as long as $g(\mathbf{0}) = \mathbf{0}$. Thus, it suffices to consider nonzero $\mathbf{y}$. Second, we claim that it suffices to prove the existence of a function $g : \mathbb{R}^N \to \mathbb{C}^m$ that satisfies both of the following properties

1. $\|g(\mathbf{y})\|_2 \leq \|\mathbf{y}\|_2$, and
2. $|\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}'\rangle\right) - \langle \mathbf{y}, \mathbf{x}'\rangle| \leq \epsilon\|\mathbf{y}\|_2$ for all $\mathbf{x}'$ in a finite $(\epsilon/2\max\{1, \|\Phi\|_{2\to 2}\})$-cover $\mathcal{C}$ of $S_{\mathcal{M}}$,

for all $\mathbf{y} \in \mathbb{R}^N$. To see why, fix $\mathbf{y} \neq \mathbf{0}$, $\mathbf{x} \in S_{\mathcal{M}}$, and let $\mathbf{x}' \in \mathcal{C} \subset S_{\mathcal{M}}$ satisfy $\|\mathbf{x}-\mathbf{x}'\|_2 \leq \epsilon/2\max\{1, \|\Phi\|_{2\to 2}\}$. We can see that any function g satisfying both of the properties above will have

$$\begin{aligned}
|\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}\rangle\right) - \langle \mathbf{y}, \mathbf{x}\rangle| &= |\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}'\rangle\right) + \Re\left(\langle g(\mathbf{y}), \Phi\left(\mathbf{x}-\mathbf{x}'\right)\rangle\right) - \langle \mathbf{y}, (\mathbf{x}-\mathbf{x}')\rangle - \langle \mathbf{y}, \mathbf{x}'\rangle| \\
&\leq |\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}'\rangle\right) - \langle \mathbf{y}, \mathbf{x}'\rangle| + |\langle g(\mathbf{y}), \Phi\left(\mathbf{x}-\mathbf{x}'\right)\rangle| + |\langle \mathbf{y}, (\mathbf{x}-\mathbf{x}')\rangle| \\
&\leq \epsilon\|\mathbf{y}\|_2 + \|g(\mathbf{y})\|_2\|\Phi\|_{2\to 2}\|\mathbf{x}-\mathbf{x}'\|_2 + \|\mathbf{y}\|_2\|\|\mathbf{x}-\mathbf{x}'\|_2
\end{aligned}$$

where the second property was used in the last inequality above.

Appealing to the first property above we can now also see that $|\Re\left(\langle g(\mathbf{y}), \Phi\mathbf{x}\rangle\right) - \langle \mathbf{y}, \mathbf{x}\rangle| \leq 2\epsilon\|\mathbf{y}\|_2$ will hold. Finally, as a consequence of the definition of $S_{\mathcal{M}}$, we therefore have that (3.3) will hold for all $\mathbf{x} \in \mathcal{M}-\mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$ whenever Properties 1 and 2 hold above. Showing that (3.3) holds all $\mathbf{x} \in \overline{\mathcal{M}}-\overline{\mathcal{M}}$ more generally can be proven by contradiction using a limiting argument combined with the fact that both the right and left hand sides of (3.3) are continuous in $\mathbf{x}$ for fixed $\mathbf{y}$. Hence, we have reduced the proof to constructing a function g that satisfies both Properties 1 and 2 above.

Let

$$g(\mathbf{y}) := \arg\min_{\mathbf{v}\in\overline{B^{2m}_{\ell^2}(\mathbf{0},\|\mathbf{y}\|_2)}} \max_{\boldsymbol{\lambda}\in\overline{B^{|\mathcal{C}|}_{\ell^1}(\mathbf{0},1)}} h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda}), \text{ where} \tag{3.4}$$

$$h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda}) := \sum_{\mathbf{u}\in\mathcal{C}} \left(\lambda_{\mathbf{u}}\left(\langle \mathbf{y}, \mathbf{u}\rangle - \Re\left(\langle \mathbf{v}, \Phi\mathbf{u}\rangle\right)\right) - \epsilon\,|\lambda_{\mathbf{u}}| \cdot \|\mathbf{y}\|_2\right) \tag{3.5}$$

where we identify $\mathbb{C}^m$ with $\mathbb{R}^{2m}$ above. Note that Property 1 above is guaranteed by definition (3.4). Furthermore, we note that if

$$\max_{\boldsymbol{\lambda}\in\{\pm\mathbf{e}_j\}_{j=1}^{|\mathcal{C}|}} h_{\mathbf{y}}(g(\mathbf{y}),\boldsymbol{\lambda}) \;=\; \max_{\mathbf{u}\in\mathcal{C}}\left(|\langle \mathbf{y}, \mathbf{u}\rangle - \Re\left(\langle g(\mathbf{y}), \Phi\mathbf{u}\rangle\right)| - \epsilon\|\mathbf{y}\|_2\right) \;\leq\; \max_{\boldsymbol{\lambda}\in\overline{B^{|\mathcal{C}|}_{\ell^1}(\mathbf{0},1)}} h_{\mathbf{y}}(g(\mathbf{y}),\boldsymbol{\lambda}) \;\leq\; 0$$

then Property 2 above will hold as well. Thus, it suffices to show that $\min_{\mathbf{v}\in\overline{B^{2m}_{\ell^2}(\mathbf{0},\|\mathbf{y}\|_2)}} \max_{\boldsymbol{\lambda}\in\overline{B^{|\mathcal{C}|}_{\ell^1}(\mathbf{0},1)}} h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda}) \leq 0$ always holds in order to finish the proof.

Noting that $h_{\mathbf{y}} : \mathbb{R}^{2m+|\mathcal{C}|} \mapsto \mathbb{R}$ defined in (3.5) is continuous, convex (affine) in $\mathbf{v}$, concave in $\boldsymbol{\lambda}$, and further noting that both $\overline{B_{\ell^1}^{|\mathcal{C}|}(\mathbf{0},1)}$ and $\overline{B_{\ell^2}^{2m}(\mathbf{0},\|\mathbf{y}\|_2)}$ are compact and convex, we may apply Von Neumann's minimax theorem [21] to see that

$$\min_{\mathbf{v}\in\overline{B_{\ell^2}^{2m}(\mathbf{0},\|\mathbf{y}\|_2)}} \max_{\boldsymbol{\lambda}\in\overline{B_{\ell^1}^{|\mathcal{C}|}(\mathbf{0},1)}} h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\in\overline{B_{\ell^1}^{|\mathcal{C}|}(\mathbf{0},1)}} \min_{\mathbf{v}\in\overline{B_{\ell^2}^{2m}(\mathbf{0},\|\mathbf{y}\|_2)}} h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda})$$

holds. Thus, we will in fact be finished if we can show that $\min_{\mathbf{v}\in\overline{B_{\ell^2}^{2m}(\mathbf{0},\|\mathbf{y}\|_2)}} h_{\mathbf{y}}(\mathbf{v},\boldsymbol{\lambda}) \leq 0$ holds for each $\boldsymbol{\lambda} \in \overline{B_{\ell^1}^{|\mathcal{C}|}(\mathbf{0},1)}$. By rescaling this in turn is implied by showing that $\forall \mathbf{u} \in \operatorname{conv}(\mathcal{C}\cup-\mathcal{C})\ \exists \mathbf{v} \in \overline{B_{\ell^2}^{2m}(\mathbf{0},\|\mathbf{y}\|_2)}$ such that

$$(\langle \mathbf{y},\mathbf{u}\rangle - \Re\left(\langle \mathbf{v},\Phi\mathbf{u}\rangle\right) - \epsilon\|\mathbf{y}\|_2) \leq 0 \tag{3.6}$$

holds.

To prove (3.6) for a fixed $\mathbf{u} \in \operatorname{conv}(\mathcal{C}\cup-\mathcal{C}) \subseteq \operatorname{conv}(S_{\mathcal{M}}\cup -S_{\mathcal{M}}) = \operatorname{conv}(S_{\mathcal{M}})$ and thereby establish the stated theorem, one may set $\mathbf{v} = \|\mathbf{y}\|_2 \frac{\Phi\mathbf{u}}{\|\Phi\mathbf{u}\|_2}$. Doing so we see that the left side of (3.6) simplifies to $\langle \mathbf{y},\mathbf{u}\rangle - \|\mathbf{y}\|_2\|\Phi\mathbf{u}\|_2 - \epsilon\|\mathbf{y}\|_2$. To finish, we note that indeed

$$\begin{aligned}\langle \mathbf{y},\mathbf{u}\rangle - \|\mathbf{y}\|_2\|\Phi\mathbf{u}\|_2 - \epsilon\|\mathbf{y}\|_2 &\leq \|\mathbf{y}\|_2\|\mathbf{u}\|_2 - \|\mathbf{y}\|_2\|\Phi\mathbf{u}\|_2 - \epsilon\|\mathbf{y}\|_2 \\ &\leq \|\mathbf{y}\|_2\left(\|\mathbf{u}\|_2 - \|\Phi\mathbf{u}\|_2 - \epsilon\right) \leq 0\end{aligned}$$

will then hold since Φ provides ϵ-convex hull distortion for $S_{\mathcal{M}}$. □

Lemma 3.4. *Let $\mathcal{M} \subset \mathbb{R}^N$ be non-empty, $\epsilon \in (0,1)$, and suppose that $\Phi \in \mathbb{C}^{m\times N}$ provides ϵ-convex hull distortion for $S_{\mathcal{M}}$. Then, there exists an outer bi-Lipschitz extension of Φ, $f : \mathbb{R}^N \to \mathbb{C}^{m+1}$, with the property that*

$$\left|\|f(\mathbf{x})-f(\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right| \leq 24\epsilon\,\|\mathbf{x}-\mathbf{y}\|_2^2 \tag{3.7}$$

holds for all $\mathbf{x} \in \mathcal{M}$ and $\mathbf{y} \in \mathbb{R}^N$.

Proof. Given $\mathbf{y} \in \mathbb{R}^N$ let $\mathbf{y}_{\mathcal{M}} \in \overline{\mathcal{M}}$ satisfy $\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2 = \inf_{\mathbf{x}\in\overline{\mathcal{M}}}\|\mathbf{y}-\mathbf{x}\|_2$.[2] We define

$$f(\mathbf{y}) := \begin{cases} (\Phi\mathbf{y},0) & \text{if } \mathbf{y}\in\overline{\mathcal{M}} \\ \left(\Phi\mathbf{y}_{\mathcal{M}} + g(\mathbf{y}-\mathbf{y}_{\mathcal{M}}), \sqrt{\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2 - \|g\left(\mathbf{y}-\mathbf{y}_{\mathcal{M}}\right)\|_2^2}\right) & \text{if } \mathbf{y}\notin\overline{\mathcal{M}} \end{cases}$$

where g is defined as in Lemma 3.3. Fix $\mathbf{x} \in \mathcal{M}$. If $\mathbf{y} \in \overline{\mathcal{M}}$ then $\|f(\mathbf{x})-f(\mathbf{y})\|_2^2 = \|\Phi(\mathbf{x}-\mathbf{y})\|_2^2$, and so we can see that $\left|\|f(\mathbf{x})-f(\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right| \leq 3\epsilon\|\mathbf{x}-\mathbf{y}\|_2^2$ will hold since Φ will be 3ϵ-JL embedding of $\overline{\mathcal{M}}-\overline{\mathcal{M}}$ (recall the proof of Corollary 3.1 and note the linearity of Φ). Thus, it suffices to consider a fixed $\mathbf{y} \notin \overline{\mathcal{M}}$. In that case we have

$$\begin{aligned}\|f(\mathbf{x})-f(\mathbf{y})\|_2^2 &= \|\Phi(\mathbf{x}-\mathbf{y}_{\mathcal{M}}) - g\left(\mathbf{y}-\mathbf{y}_{\mathcal{M}}\right)\|_2^2 + \|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2 - \|g\left(\mathbf{y}-\mathbf{y}_{\mathcal{M}}\right)\|_2^2 \\ &= \|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2 + \|\Phi(\mathbf{x}-\mathbf{y}_{\mathcal{M}})\|_2^2 - 2\Re\left(\langle g\left(\mathbf{y}-\mathbf{y}_{\mathcal{M}}\right), \Phi(\mathbf{x}-\mathbf{y}_{\mathcal{M}})\rangle\right)\end{aligned} \tag{3.8}$$

by the polarization identity and parallelogram law.

Similarly we have that

$$\|\mathbf{x}-\mathbf{y}\|_2^2 = \|(\mathbf{x}-\mathbf{y}_{\mathcal{M}}) - (\mathbf{y}-\mathbf{y}_{\mathcal{M}})\|_2^2 = \|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2 + \|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2 - 2\langle \mathbf{y}-\mathbf{y}_{\mathcal{M}}, \mathbf{x}-\mathbf{y}_{\mathcal{M}}\rangle. \tag{3.9}$$

Subtracting (3.9) from (3.8) we can now see that

$$\begin{aligned}\left|\|f(\mathbf{x})-f(\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right| \leq\ & \left|\|\Phi(\mathbf{x}-\mathbf{y}_{\mathcal{M}})\|_2^2 - \|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2\right| + \\ & 2\left|\Re\left(\langle g\left(\mathbf{y}-\mathbf{y}_{\mathcal{M}}\right), \Phi(\mathbf{x}-\mathbf{y}_{\mathcal{M}})\rangle\right) - \langle \mathbf{y}-\mathbf{y}_{\mathcal{M}}, \mathbf{x}-\mathbf{y}_{\mathcal{M}}\rangle\right|\end{aligned}$$

[2] One can see that it suffices to approximately compute $\mathbf{y}_{\mathcal{M}}$ in order to achieve (3.7) up to a fixed precision.

$$\begin{aligned} &\leq 3\epsilon\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2 \;+\; 4\epsilon\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2 \\ &\leq 3\epsilon\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2 \;+\; 2\epsilon\left(\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2+\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2\right) \end{aligned} \tag{3.10}$$

where the second inequality again appeals to Φ being a 3ϵ-JL embedding of $\overline{\mathcal{M}}-\overline{\mathcal{M}}$, and to Lemma 3.3. Considering (3.10) we can see that

- $\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2 \leq \|\mathbf{y}-\mathbf{x}\|_2$ by the definition of $\mathbf{y}_{\mathcal{M}}$, and so
- $\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2+\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2 \leq 2\|\mathbf{x}-\mathbf{y}\|_2$, and thus
- $\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2^2+\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2^2 \leq \left(\|\mathbf{y}-\mathbf{y}_{\mathcal{M}}\|_2+\|\mathbf{x}-\mathbf{y}_{\mathcal{M}}\|_2\right)^2 \leq 9\|\mathbf{x}-\mathbf{y}\|_2^2$.

Using the last two inequalities above in (3.10) now yields the stated result. □

We are now prepared to prove the two main results of this section.

3.4 Proof of Theorem 3.1

Apply Theorem 3.4 with $\epsilon \leftarrow \epsilon/24$ in order obtain $\epsilon/24$-convex hull distortion for $S_{\mathcal{M}}$ via Φ. Then, apply Lemma 3.4.

3.5 Proof of Theorem 3.2

To begin we apply Corollary 3.1 with, e.g., $p=1/2$ to demonstrate that an $\left\lceil \frac{c''}{\epsilon^2}\left(w(S_{\mathcal{M}})+\sqrt{\ln(4)}\right)^2\right\rceil \times N$ matrix with i.i.d. standard normal random entries can provide $(\epsilon/24)$-convex hull distortion for $S_{\mathcal{M}}$, where c'' is an absolute constant. Hence, such a matrix Φ exists. An application of Lemma 3.4 now finishes the proof.

4 The Proof of Theorem 1.1

We apply Theorem 3.2 together with Theorem 2.1 to bound the Gaussian width of $S_{\mathcal{M}}$.

5 A Numerical Evaluation of Terminal Embeddings

In this section we consider several variants of the optimization approach mentioned in Section 3.3 of [19] for implementing a terminal embedding $f:\mathbb{R}^N \to \mathbb{R}^{m+1}$ of a finite set $X \subset \mathbb{R}^N$. In effect, this requires us to implement a function satisfying two sets of constraints from [19, Section 3.3] that are analogous to the two properties of $g:\mathbb{R}^N \to \mathbb{C}^m$ listed at the beginning of the proof of Lemma 3.3. See Lines 1 and 2 of Algorithm 1 for a concrete example of one type of constrained minimization problem solved herein to accomplish this task.

Algorithm 1 Terminal Embedding of a Finite Set

Input: $\epsilon \in (0,1)$, $X \subset \mathbb{R}^N$, $|X| =: n$, $S \subset \mathbb{R}^N$, $|S| =: n'$, $m \in \mathbb{N}$ with $m < N$, a random matrix with i.i.d. standard Gaussian entries, $\Phi \in \mathbb{R}^{m \times N}$, rescaled to perform as a JL embedding matrix $\Pi := \frac{1}{\sqrt{m}}\Phi$

Output: A terminal embedding of X, $f \in \mathbb{R}^N \to \mathbb{R}^{m+1}$, evaluated on S

for $\mathbf{u} \in S$ **do**

1) Compute $\mathbf{x}_{NN} := \text{argmin}_{\mathbf{x} \in X} \|\mathbf{u} - \mathbf{x}\|_2$

2) Solve the following constrained minimization problem to compute a minimizer $\mathbf{u}' \in \mathbb{R}^m$

$$\begin{aligned} \text{Minimize} \quad & h_{\mathbf{u},\mathbf{x}_{NN}}(\mathbf{z}) := \|\mathbf{z}\|_2^2 + 2\langle \Pi(\mathbf{u} - \mathbf{x}_{NN}), \mathbf{z} \rangle \\ \text{subject to} \quad & \|\mathbf{z}\|_2 \leq \|\mathbf{u} - \mathbf{x}_{NN}\|_2 \\ & |\langle \mathbf{z}, \Pi(\mathbf{x} - \mathbf{x}_{NN}) \rangle - \langle \mathbf{u} - \mathbf{x}_{NN}, \mathbf{x} - \mathbf{x}_{NN} \rangle| \leq \epsilon \|\mathbf{u} - \mathbf{x}_{NN}\|_2 \|\mathbf{x} - \mathbf{x}_{NN}\|_2 \ \forall \mathbf{x} \in X \end{aligned}$$

3) Compute $f : \mathbb{R}^N \to \mathbb{R}^{m+1}$ at $\mathbf{u}$ via

$$f(\mathbf{u}) := \begin{cases} (\Pi\mathbf{u}, 0), & \mathbf{u} \in X \\ (\Pi\mathbf{x}_{NN} + \mathbf{u}', \sqrt{\|\mathbf{u} - \mathbf{x}_{NN}\|_2^2 - \|\mathbf{u}'\|_2^2}), & \mathbf{u} \notin X \end{cases}$$

end for

Crucially, we note that any choice $\mathbf{u}' \in \mathbb{R}^m$ of a $\mathbf{z}$ satisfying the two sets of constraints in Line 2 of Algorithm 1 for a given $\mathbf{u} \in \mathbb{R}^N$ is guaranteed to correspond to an evaluation of a valid terminal embedding of X at $\mathbf{u}$ in Line 3. This leaves the choice of the objective function, $h_{\mathbf{u},\mathbf{x}_{NN}}$, minimized in Line 2 of Algorithm 1 open to change without effecting its theoretical performance guarantees. Given this setup, several heretofore unexplored practical questions about terminal embeddings immediately present themselves. These include:

1. Repeatedly solving the optimization problem in Line 2 of Algorithm 1 to evaluate a terminal embedding of X on S is certainly more computationally expensive than simply evaluating a standard linear Johnson-Lindenstrauss (JL) embedding of X on S instead. How do terminal embeddings empirically compare to standard linear JL embedding matrices on real-world data in the context of, e.g., compressive classification? When, if ever, is their additional computational expense actually justified in practice?

2. Though any choice of objective function $h_{\mathbf{u},\mathbf{x}_{NN}}$ in Line 2 of Algorithm 1 must result in a terminal embedding f of X based on the available theory, some choices probably lead to better empirical performance than others. What's a good default choice?

3. How much dimensionality reduction are terminal embeddings capable of in the context of, e.g., accurate compressive classification using real-world data?

In keeping with the motivating application discussed in Section 1.1 above, we will explore some preliminary answers to these three questions in the context of compressive classification based on real-world data below.

5.1 A Comparison Criteria: Compressive Nearest Neighbor Classification

Given a labelled data set $\mathcal{D} \subset \mathbb{R}^N$ with label set $\mathcal{L}$, we let $Label : \mathcal{D} \to \mathcal{L}$ denote the function which assigns the correct label to each element of the data set. To address the three questions above we will use compressive nearest neighbor classification accuracy as a primary measure of an embedding strategy's quality. See Algorithm 2 for a detailed description of how this accuracy can be computed for a given data set $\mathcal{D}$.

Algorithm 2 Measuring Compressive Nearest Neighbor Classification Accuracy

Input: $\epsilon \in (0,1)$, A labeled data set $\mathcal{D} \subset \mathbb{R}^N$ split into two disjoint subsets: A training set $X \subset \mathcal{D}$ with $|X| =: n$, and a test set $S \subset \mathcal{D}$ with $|S| =: n'$, such that $S \cap X = \emptyset$. A compressive dimension $m < N$.

Output: Successful Nearest Neighbor Classification Percentage for Data Embedded in $\mathbb{R}^{m+1}$

Fix $f : \mathbb{R}^N \to \mathbb{R}^{m+1}$, an embedding of the training data $X \subset \mathbb{R}^N$ into $\mathbb{R}^{m+1}$ satisfying

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|f(\mathbf{x})-f(\mathbf{y})\|_2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2$$

for all $\mathbf{x}, \mathbf{y} \in X$. [Note: this can either be a JL-embedding of X, or a stronger terminal embedding of X.]

% *Embed the training data into* $\mathbb{R}^{m+1}$.
for $\mathbf{x} \in X$ **do**
  Compute $f(\mathbf{x})$ using, e.g., Algorithm 1.
end for

% *Classify the test data using its embedded distance in* $\mathbb{R}^{m+1}$.
$p = 0$
for $\mathbf{u} \in S$ **do**
  Compute $f(\mathbf{u})$ using, e.g., Algorithm 1
  Compute $\mathbf{x} = \text{argmin}_{\mathbf{y} \in X} \|f(\mathbf{u}) - f(\mathbf{y})\|_2$
  if $Label(\mathbf{u}) = Label(\mathbf{x})$ **then**
    $p = p + 1$
  end if
end for

Output the Successful Classification Percentage $= \dfrac{p}{n'} \times 100\%$

Note that Algorithm 2 can be used to help us compare the quality of different embedding strategies. For example, one can use Algorithm 2 to compare different choices of objective functions $h_{\mathbf{u},\mathbf{x}_{NN}}$ in Line 2 of Algorithm 1 against one another by running Algorithm 2 multiple times on the same training and test data sets while only varying the implementation of Algorithm 1 each time. This is exactly the type of approach we will use below. Of course, before we can begin we must first decide on some labelled data sets $\mathcal{D}$ to use in our classification experiments.

5.2 Our Choice of Training and Testing Data Sets

Herein we consider two standard benchmark image data sets which allow for accurate uncompressed Nearest Neighbor (NN) classification. The images in each data set can then be vectorized and embedded using, e.g., Algorithm 1 in order to test the accuracies of compressed NN classification variants against both one another, as well as against standard uncompressed NN classification. These benchmark data sets are as follows.

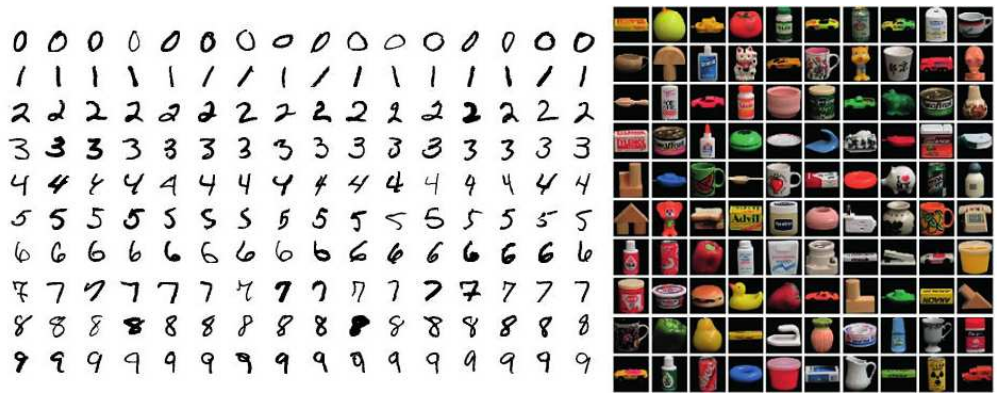

Figure 1: Example images from the MNIST data set (left), and the COIL-100 data set (right).

The MNIST data set [17, 6] consists of 60,0000 training images of 28×28-pixel grayscale hand-written images of the digits 0 through 9. Thus, MNIST 10 labels to correctly classify between, and $N = 28^2 = 784$. For all experiments involving the MNIST dataset $n/10$ digits of each type are selected uniformly at random to form the training set X, for a total of n vectorized training images in $\mathbb{R}^{784}$. Then, 100 digits of each type are randomly selected from those not used for training in order to form the test set S, leading to a total of $n' = 1000$ vectorized test images in $\mathbb{R}^{784}$. See the left side of Figure 1 for example MNIST images.

The COIL-100 data set [20] is a collection of 128×128-pixel color images of 100 objects, each photographed 72 times where the object has been rotated by 5 degrees each time to get a complete rotation. However, only the green color channel of each image is used herein for simplicity. Thus, herein COIL-100 consists of $7,200$ total vectorized images in $\mathbb{R}^N$ with $N = 128^2 = 16,384$, where each image has one of 100 different labels (72 images per label). For all experiments involving this COIL-100 data set, $n/100$ training images are down sampled from each of the 100 objects' rotational image sequences. Thus, the training sets each contain $n/100$ vectorized images of each object, each photographed at rotations of $\approx 36000/n$ degrees (rounded to multiples of 5). The resulting training data sets therefore all consist of n vectorized images in $\mathbb{R}^{16,384}$. After forming each training set, 10 images of each type are then randomly selected from those not used for training in order to form the test set S, leading to a total of $n' = 1000$ vectorized test images in $\mathbb{R}^{16,384}$ per experiment. See the right side of Figure 1 for example COIL-100 images.

5.3 A Comparison of Four Embedding Strategies via NN Classification

In this section we seek to better understand (i) when terminal embeddings outperform standard JL-embedding matrices in practice with respect to accurate compressive NN classification, (ii) what type of objective functions $h_{\mathbf{u},\mathbf{x}_{NN}}$ in Line 2 of Algorithm 1 perform best in practice when computing a terminal embedding, and (iii) how much dimensionality reduction one can achieve with a terminal embedding without appreciably degrading standard NN classification results in practice. To gain insight on these three questions we will compare the following four embedding strategies in the context of NN classification. These strategies begin with the most trivial linear embeddings (i.e., the identity map) and slowly progress toward extremely non-linear terminal embeddings.

(a) **Identity:** We use the data in its original uncompressed form (i.e., we use the trivial embedding $f : \mathbb{R}^N \to \mathbb{R}^N$ defined by $f(\mathbf{u}) = \mathbf{u}$ in Algorithm 2). Here the embedding dimension $m + 1$ is always fixed to be N.

(b) **Linear:** We compressively embed our training data X using a JL embedding. More specifically, we generate an $m \times N$ random matrix Φ with i.i.d. standard Gaussian entries and then set $f : \mathbb{R}^N \to \mathbb{R}^{m+1}$ to be $f(\mathbf{u}) := \left(\frac{1}{\sqrt{m}}\Phi\mathbf{u},\ 0\right)$ in Algorithm 2 for various choices of m. It is then hoped that f will embed

the test data S well in addition to the training data X. Note that this embedding choice for f is consistent with Algorithm 1 where one lets $X = X \cup S$ when evaluating Line 3, thereby rendering the minimization problem in Line 2 irrelevant.

(c) **A Valid Terminal Embedding That's as Linear as Possible:** To minimize the pointwise difference between the terminal embedding f computed by Algorithm 1 and the linear map defined above in (b), we may choose the objective function in Line 2 of Algorithm 1 to be $h_{\mathbf{u},\mathbf{x}_{NN}}(\mathbf{z}) := \langle \Pi(\mathbf{x}_{NN} - \mathbf{u}), \mathbf{z} \rangle$. To see why solving this minimizes the pointwise difference between f and the linear map in (b), let $\mathbf{u}'$ be such that $\langle \Pi(\mathbf{x}_{NN} - \mathbf{u}), \mathbf{z} \rangle$ is minimal subject to the constraints in Line 2 of Algorithm 1 when $\mathbf{z} = \mathbf{u}'$. Since $\mathbf{u}$ and $\mathbf{x}_{NN}$ are fixed here, we note that $\mathbf{z} = \mathbf{u}'$ will then also minimize

$$\begin{aligned} & \|\Pi(\mathbf{x}_{NN} - \mathbf{u})\|_2^2 + 2\langle \Pi(\mathbf{x}_{NN} - \mathbf{u}), \mathbf{z} \rangle + \|\mathbf{u} - \mathbf{x}_{NN}\|_2^2 \\ = \; & \|\Pi(\mathbf{x}_{NN} - \mathbf{u})\|_2^2 + \|\mathbf{z}\|_2^2 + 2\langle \Pi(\mathbf{x}_{NN} - \mathbf{u}), \mathbf{z} \rangle + \|\mathbf{u} - \mathbf{x}_{NN}\|_2^2 - \|\mathbf{z}\|_2^2 \\ = \; & \|\Pi(\mathbf{x}_{NN} - \mathbf{u}) + \mathbf{z}\|_2^2 + \|\mathbf{u} - \mathbf{x}_{NN}\|_2^2 - \|\mathbf{z}\|_2^2 \\ = \; & \left\| \left(\Pi \mathbf{x}_{NN} + \mathbf{z}, \sqrt{\|\mathbf{u} - \mathbf{x}_{NN}\|_2^2 - \|\mathbf{z}\|_2^2} \right) - (\Pi \mathbf{u}, 0) \right\|_2^2 \end{aligned}$$

subject to the desired constraints. Hence, we can see that choosing $\mathbf{z} = \mathbf{u}'$ as above is equivalent to minimizing $\|f(\mathbf{u}) - (\Pi \mathbf{u}, 0)\|_2^2$ over all valid choices of terminal embeddings f that satisfy the existing theory.

(d) **A Terminal Embedding Computed by Algorithm 1 as Presented:** This terminal embedding is computed using Algorithm 1 exactly as it is formulated above (i.e., with the objective function in Line 2 chosen to be $h_{\mathbf{u},\mathbf{x}_{NN}}(\mathbf{z}) := \|\mathbf{z}\|_2^2 + 2\langle \Pi(\mathbf{u} - \mathbf{x}_{NN}), \mathbf{z} \rangle$). Note that this choice of objective function was made to encourage non-linearity in the resulting terminal embedding f computed by Algorithm 1. To understand our intuition for making this choice of objective function in order to encourage non-linearity in f, suppose that $\|\mathbf{z}\|_2^2 + 2\langle \Pi(\mathbf{u} - \mathbf{x}_{NN}), \mathbf{z} \rangle$ is minimal subject to the constraints in Line 2 of Algorithm 1 when $\mathbf{z} = \mathbf{u}'$. Since $\mathbf{u}$ and $\mathbf{x}_{NN}$ are fixed independently of $\mathbf{z}$ this means that $\mathbf{z} = \mathbf{u}'$ then also minimize

$$\|\mathbf{z}\|_2^2 + 2\langle \Pi(\mathbf{u} - \mathbf{x}_{NN}), \mathbf{z} \rangle + \|\Pi(\mathbf{u} - \mathbf{x}_{NN})\|_2^2 = \|\mathbf{z} + \Pi(\mathbf{u} - \mathbf{x}_{NN})\|_2^2.$$

Hence, this objection function is encouraging $\mathbf{u}'$ to be as close to $-\Pi(\mathbf{u} - \mathbf{x}_{NN}) = \Pi(\mathbf{x}_{NN} - \mathbf{u})$ as possible subject to satisfying the constraints in Line 2 of Algorithm 1. Recalling (c) just above, we can now see that this is exactly encouraging $\mathbf{u}'$ to be a value for which the objective function we seek to minimize in (c) is relatively large.

We are now prepared to empirically compare the four types of embeddings (a) – (d) on the data sets discussed above in Section 5.2. To do so, we run Algorithm 2 four times for several different choices of embedding dimension m on each data set below, varying the choice of embedding f between (a), (b), (c), and (d) for each value of m. The successful classification percentage is then plotted as a function of m for each different data set and choice of embedding. See Figures 2(a) and 2(c) for the results. In addition, to quantify the extent to which the embedding strategies (b) – (d) above are increasingly nonlinear, we also measure the relative distance between where each training-set embedding f maps points in the test sets versus where its associated linear training-set embedding would map them. More specifically, for each embedding f and test point $\mathbf{u} \in S$ we let

$$\text{Nonlinearity}_f(\mathbf{u}) = \frac{\|f(\mathbf{u}) - (\Pi \mathbf{u}, 0)\|_2}{\|(\Pi \mathbf{u}, 0)\|_2} \times 100\%$$

See Figures 2(b) and 2(d) for plots of

$$\text{Mean}_{\mathbf{u} \in S} \; \text{Nonlinearity}_f(\mathbf{u})$$

for each of the embedding strategies (b) – (d) on the data sets discussed in Section 5.2.

To compute solutions to the minimization problem in Line 2 of Algorithm 1 below we used the MATLAB package CVX [10, 9] with the initialization $\mathbf{z}_0 = \Pi(\mathbf{u} - \mathbf{x}_{NN})$ and $\epsilon = 0.1$ in the constraints. All simulations were performed using MATLAB R2021b on an Intel desktop with a 2.60GHz i7-10750H CPU and 16GB DDR4 2933MHz memory. All code used to generate the figures below is publicly available at `https://github.com/MarkPhilipRoach/TerminalEmbedding`.

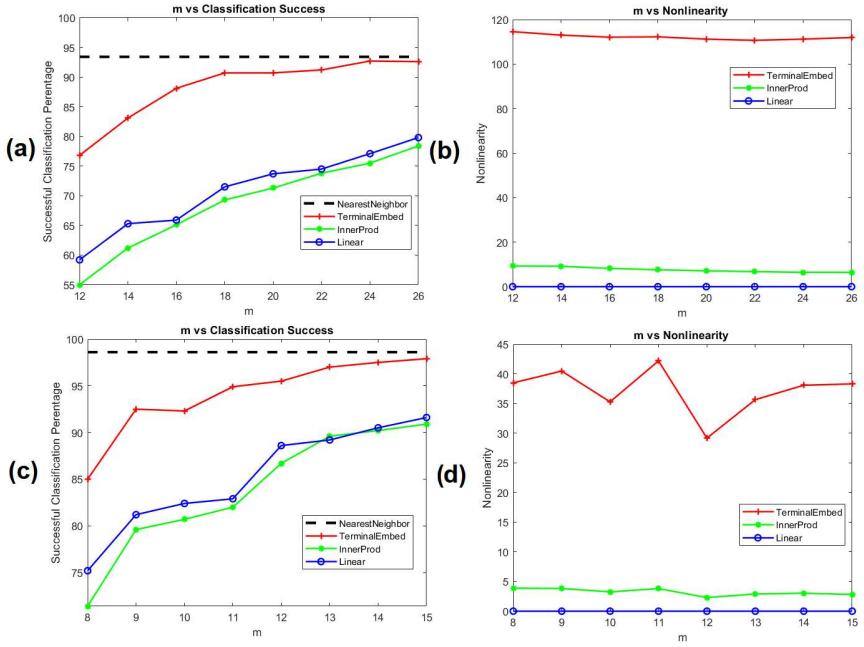

Figure 2: Figures 2(a) and 2(b) concern the MNIST data set with training set size $n = 4000$ and test set size $n' = 1000$ in all experiments. Similarly, Figures 2(c) and 2(d) concern the COIL-100 dataset with training set size $n = 3600$ and test set size $n' = 1000$ in all experiments. In both Figures 2(a) and 2(c) the dashed black "NearestNeighbor" line plots the classification accuracy when the **Identity** map (a) is used in Algorithm 2. Note that the "NearestNeighbor" line is independent of m because the indentity map involves no compression. Similarly, in all of the Figures 2(a) – 2(d) the red "TerminalEmbed" curves correspond to the use of Algorithm 1 as it's presented to compute highly non-linear terminal embeddings (embedding strategy (d) above), the green "InnerProd" curves correspond to the use of nearly linear terminal embeddings (embedding strategy (c) above), and the blue "Linear" curves correspond to the use of **Linear** JL embedding matrices (embedding strategy (b) above).

Looking at Figure 2 one can see that the most non-linear embedding strategy (d) – i.e., Algorithm 1 – allows for the best compressed NN classification performance, outperforming standard linear JL embeddings for all choices of m. Perhaps most interestingly, it also quickly converges to the uncompressed NN classification performance, matching it to within 1 percent at the values of $m = 24$ for MNIST and $m = 15$ for COIL-100. This corresponds to relative dimensionality reductions of

$$100(1 - 24/784)\% \approx 96.9\%$$

and

$$100(1 - 15/16384)\% \approx 99.9\%,$$

respectively, with negligible loss of NN classification accuracy. As a result, it does indeed appear as if nonlinear terminal embeddings have the potential to allow for improvements in dimensionality reduction in the context of classification beyond what standard linear JL embeddings can achieve.

Of course, challenges remain in the practical application of such nonlinear terminal embeddings. Principally, their computation by, e.g., Algorithm 1 is orders of magnitude slower than simply applying a JL embedding matrix to the data one wishes to compressively classify. Nonetheless, if dimension reduction at all costs is one's goal, terminal embeddings appear capable of providing better results than their linear brethren. And, recent theoretical work [3] aimed at lessening their computational deficiencies looks promising.

5.4 Additional Experiments on Effective Distortions and Run Times

In this section we further investigate the best performing terminal embedding strategy from the previous section (i.e., Algorithm 1) on the MNIST and COIL-100 data sets. In particular, we provide illustrative experiments concerning the improvement of (i) compressive classification accuracy with training set size, and (ii) the effective distortion of the terminal embedding with embedding dimension $m+1$. Furthermore, we also investigate (iii) the run time scaling of Algorithm 1.

To compute the effective distortions of a given (terminal) embedding of training data X, $f : \mathbb{R}^N \to \mathbb{R}^{m+1}$, over all available test and train data $X \cup S$ we use

$$\text{MaxDist}_f = \max_{\mathbf{x}\in X} \max_{\mathbf{u}\in S\cup X\setminus\{\mathbf{x}\}} \frac{\|f(\mathbf{u}) - f(\mathbf{x})\|_2}{\|\mathbf{u}-\mathbf{x}\|_2}, \quad \text{MinDist}_f = \min_{\mathbf{x}\in X} \min_{\mathbf{u}\in S\cup X\setminus\{\mathbf{x}\}} \frac{\|f(\mathbf{u}) - f(\mathbf{x})\|_2}{\|\mathbf{u}-\mathbf{x}\|_2}.$$

Note that these correspond to estimates of the upper and lower multiplicative distortions, respectively, of a given terminal embedding in (1.1). In order to better understand the effect of the minimizer $\mathbf{u}'$ of the minimization problem in Line 2 of Algorithm 1 on the final embedding f, we will also separately consider the effective distortions of its component linear JL embedding $\mathbf{u} \mapsto (\Pi\mathbf{u}, 0)$ below. See Figures 3 and 4 for such plots using the MNIST and COIL-100 data sets, respectively.

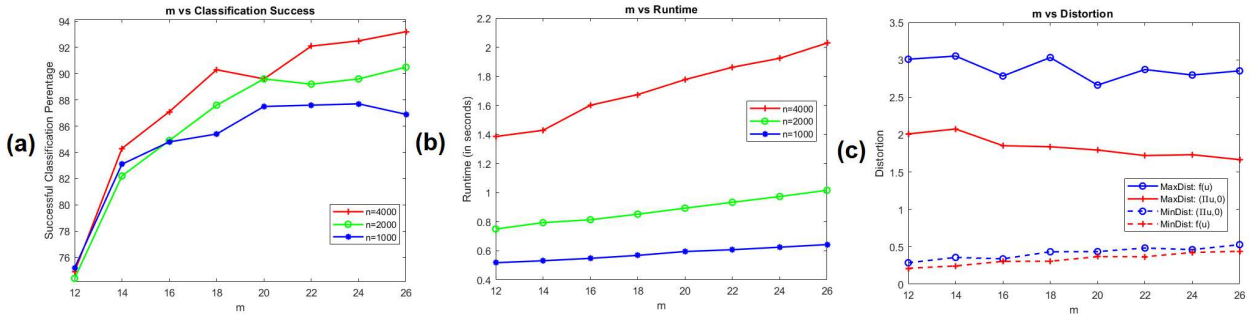

Figure 3: This figure compares (a) compressive NN classification accuracies, and (b) the classification run times of Algorithm 2 averaged over all $\mathbf{u} \in S$, on the MNIST data set. Three different training data set sizes $n = |X| \in \{1000,\ 2000,\ 4000\}$ were fixed as the embedding dimension $m+1$ varied for each of the first two subfigures. Recall that the test set size is always fixed to $n' = 1000$. In addition, Figure (c) compares MaxDist_f and MinDist_f for the nonlinear f computed by Algorithm 1 versus its component linear embedding $\mathbf{u} \mapsto (\Pi\mathbf{u}, 0)$ as m varies for a fixed embedded training set size of $n = 4000$.

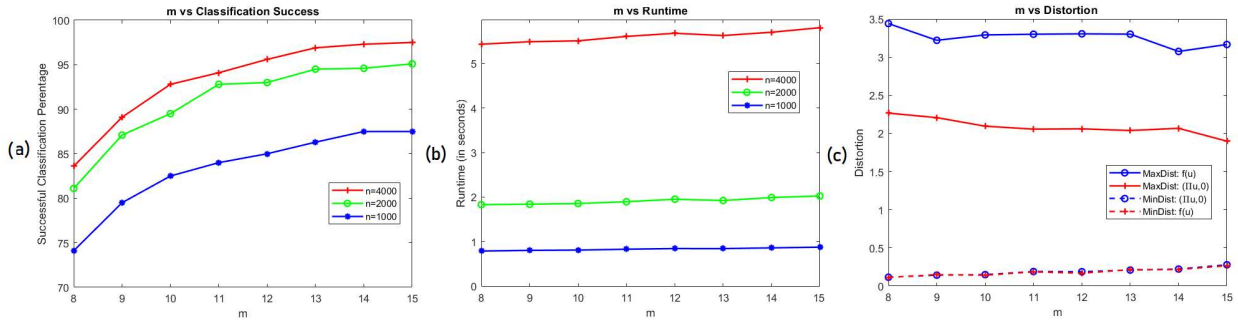

Figure 4: Figures (a) and (b) here are run with identical parameters as for their corresponding subfigures in Figure 3, except using the COIL-100 data set. Similarly, Figure (c) compares MaxDist_f and MinDist_f for the nonlinear f computed by Algorithm 1 versus its component linear embedding $\mathbf{u} \mapsto (\Pi\mathbf{u}, 0)$ as m varies for a fixed embedded training set size of $n = 3600$.

Looking at Figures 3 and 4 one notes several consistent trends. First, compressive classification accuracy increases with both training set size n and embedding dimension m, as generally expected. Second, compressive classification run times also increase with training set size n (as well as more mildly with embedding dimension m). This is mainly due to the increase in the number of constraints in Line 2 of Algorithm 1 with the training set size n. Finally, the distortion plots indicate that the nonlinear terminal embeddings f computed by Algorithm 1 tend to preserve the lower distortions of their component linear JL embeddings while simultaneously increasing their upper distortions. As a result, the nonlinear terminal embeddings considered here appear to spread the initially JL embedded data out, perhaps pushing different classes away from one another in the process. If so, it would help explained the increased compressive NN classification accuracy observed for Algorithm 1 in Figure 2.

Acknowledgements

Mark Iwen was supported in part by NSF DMS 2106472. Mark Philip Roach was supported in part by NSF DMS 1912706. Mark Iwen would like to dedicate his effort on this paper to William E. Iwen, 1/27/1939 – 12/10/2021, as well as to all those at the VA Medical Center in St. Cloud, MN who cared for him so tirelessly during the COVID-19 pandemic.[3] Thank you.

References

[1] Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[2] Imre Bárány. A generalization of carathéodory's theorem. *Discrete Mathematics*, 40(2):141–152, 1982.

[3] Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1209–1216. IEEE, 2022.

[4] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[5] Mark A Davenport, Marco F Duarte, Michael B Wakin, Jason N Laska, Dharmpal Takhar, Kevin F Kelly, and Richard G Baraniuk. The smashed filter for compressive classification and target recognition.

[3] https://www.greenbaypressgazette.com/obituaries/wis343656

In *Computational Imaging V*, volume 6498, page 64980H. International Society for Optics and Photonics, 2007.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[7] Armin Eftekhari and Michael B Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.

[8] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–418, March 1959.

[9] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[10] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[11] Mark Iwen, Arman Tavakoli, and Benjamin Schmidt. Lower bounds on the low-distortion embedding dimension of submanifolds of $\mathbb{R}^n$. *arXiv preprint arXiv:2105.13512*, 2021.

[12] Mark A Iwen, Deanna Needell, Elizaveta Rebrova, and Ali Zare. Lower memory oblivious (tensor) subspace embeddings with fewer random bits: modewise methods for least squares. *SIAM Journal on Matrix Analysis and Applications*, 42(1):376–416, 2021.

[13] Mark A. Iwen, Benjamin Schmidt, and Arman Tavakoli. On fast johnson-lindenstrauss embeddings of compact submanifolds of R^N with boundary. *Arxiv*, 2110.04193, 2021.

[14] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

[15] Mojzesz Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934.

[16] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

[17] Y. LeCun, C. Cortes, and C.J.C. Burges. The mnist database of handwritten digits. 1998.

[18] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1088–1101, 2018.

[19] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1064–1069, 2019.

[20] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). 1996.

[21] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[22] Christoph Thäle. 50 years sets with positive reach–a survey. *Surveys in Mathematics and its Applications*, 3:123–165, 2008. Publisher: University Constantin Brancusi.

[23] Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*. Number 47 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, New York, NY, 2018.