SCALE-OBLIVIOUS METRIC FRAGMENTATION AND THE NONLINEAR DVORETZKY THEOREM

ASSAF NAOR AND TERENCE TAO

Abstract. We introduce a randomized iterative fragmentation procedure for finite metric spaces, which is guaranteed to result in a polynomially large subset that is D-equivalent to an ultrametric, where $D \in (2,\infty)$ is a prescribed target distortion. Since this procedure works for D arbitrarily close to the nonlinear Dvoretzky phase transition at distortion 2, we thus obtain a much simpler probabilistic proof of the main result of [3], answering a question from [12], and yielding the best known bounds in the nonlinear Dvoretzky theorem.

Our method utilizes a sequence of random scales at which a given metric space is fragmented. As in many previous randomized arguments in embedding theory, these scales are chosen irrespective of the geometry of the metric space in question. We show that our bounds are sharp if one utilizes such a "scale-oblivious" fragmentation procedure.

1. Introduction

A metric space (X,d) is said to embed into Hilbert space with *distortion* $D \geqslant 1$ if there exists $f : X \to \ell_2$ satisfying $d(x,y) \leqslant \|f(x)-f(y)\|_2 \leqslant Dd(x,y)$ for all $x,y \in X$. Dvoretzky's theorem [9] asserts that for every $k \in \mathbb{N}$ and $D > 1$ there exists $n = n(k,D) \in \mathbb{N}$ such that every n-dimensional normed space has a k-dimensional linear subspace that embeds into Hilbert space with distortion D; see [14, 13, 15] for the best known bounds on $n(k,D)$.

Motivated by a possible analogue of Dvoretzky's theorem in the class of general metric spaces, Bourgain, Figiel and Milman introduced in [6] the *nonlinear Dvoretzky problem*, which asks for the largest integer $k = k(n,D)$ such that any n-point metric space has a subset of cardinality k that embeds into Hilbert space with distortion D. They showed [6] that for every $D > 1$ we have $\lim_{n\to\infty} k(n,D) = \infty$, thus establishing the validity of a nonlinear Dvorezky phenomenon. Quantitatively, the main result of [6] asserts that $k(n,D) \geqslant c(D)\log n$, and that there exists $D_0 > 1$ for which $k(n,D_0) = O(\log n)$.

Renewed interest in the nonlinear Dvorezky problem due to the discovery of applications to the theory of online algorithms resulted in a sequence of works [11, 5, 1] which culminated in the following threshold phenomenon from [3] (see also [2, 4, 8] for related results):

Theorem 1.1 ([3]). *For $D > 1$ there exist $a(D), A(D) \in (0,\infty)$ and $b(D), B(D) \in (0,1)$ with the following properties:*

(1) If $D \in (1,2)$ then any n-point metric space has a subset of cardinality $\geqslant a(D)\log n$ that embeds with distortion D into Hilbert space. On the other hand, there exist arbitrarily large n-point metric spaces X_n with the property that any $Y \subseteq X_n$ that embeds into Hilbert space with distortion D necessarily satisfies $|Y| \leqslant A(D)\log n$.

(2) If $D \in (2,\infty)$ then any n-point metric space has a subset of cardinality $\geqslant n^{1-b(D)}$ that embeds with distortion D into Hilbert space. On the other hand, there exist arbitrarily large n-point metric spaces X_n with the property that any $Y \subseteq X_n$ that embeds into Hilbert space with distortion D necessarily satisfies $|Y| \leqslant n^{1-B(D)}$.

Note that the first assertion of part (1) of Theorem 1.1 is just a restatement of the Bourgain-Figiel-Milman nonlinear Dvoretzky theorem [6].

All the positive embedding results quoted above are actually stronger than embeddings into Hilbert space: they produce subsets which embed with distortion D into an *ultrametric*. Recall that a metric space (U, ρ) is an ultrametric if $\rho(u, v) \leqslant \max\{\rho(u, w), \rho(w, v)\}$ for every $u, v, w \in U$. Separable ultrametrics embed isometrically into Hilbert space [17], so the problem of finding a subset of a metric space which embeds with distortion D into an ultrametric is a (strictly) stronger statement than the nonlinear Dvoretzky problem. The fact that the embeddings of Theorem 1.1 are into ultrametrics is crucial for its applications. Note, however, that the impossibility results in Theorem 1.1 rule out embeddings into Hilbert space, and not just embeddings into ultrametrics.

The proofs of the nonlinear Dvoretzky theorems in [6, 11, 5, 1, 2, 4, 8] proceed via *deterministic* constructions. In [12] a new approach to the nonlinear Dvoretzky problem was introduced, based on a *probabilistic* argument which is closer in spirit to the proofs of the classical Dvoretzky theorem. This randomized approach, called the method of *Ramsey partitions*, has three main advantages. First, it leads to new algorithmic applications of the nonlinear Dvoretzky theorem which are very different from the applications in [11, 5, 1, 3]; we shall briefly describe one of these applications in Section 1.1. Second, Ramsey partitions yield a major simplification of the proof of part (2) in Theorem 1.1 for sufficiently large values of D (this part of Theorem 1.1 is by far the most complicated part of its proof in [3]). Third, the bound on the exponent $b(D)$ obtained in [12] is asymptotically sharp as $D \to \infty$, unlike the bound in [3], which is off by a logarithmic factor. Specifically, [12] yields $b(D) \lesssim 1/D$, which is optimal up to the implied universal constant due to the bound $B(D) \gtrsim 1/D$ of [3].

An obvious question, raised in [12], suggests itself: can the randomized approach of [12] yield a proof Theorem 1.1 in which the target distortion $D > 2$ is allowed to go all the way down to the phase transition at 2? The main result of [12] states that for $D > 2$, any n-point metric space X has a subset $Y \subseteq X$ with $|Y| \geqslant n^{1-128/D}$ which embed with distortion D into an ultrametric. [12] did not attempt to optimize the constant 128 in this result, and indeed by a more careful analysis of the arguments of [12] one can ensure that $|Y| \geqslant n^{1-16/(D-2)}$ (even this estimate can be slightly improved, but not by much). In any case, it is clear that these statements become vacuous for D smaller than a universal constant close enough to 2. Thus, the full $D > 2$ range of part (2) of Theorem 1.1 still required the use of the deterministic approach of [3].

It was stated in [12] that there does not seem to be a simple way to use Ramsey partitions to handle distortions arbitrarily close to 2. In Section 1.1 we make a very simple observation which proves that if $D < 3$, then the method of Ramsey partitions cannot yield a subset Y as above of size tending to ∞ with n. Thus, in fact, it is *impossible* to approach the phase transition at 2 using Ramsey partitions. Here we present a new randomized approach, building on the multiplicative telescoping argument of [12], which proves the nonlinear Dvoretzky theorem for any distortion $D > 2$. Specifically, we prove the following result:

Theorem 1.2. *For every $D > 2$, any n-point metric space has a subset of cardinality $n^{\theta(D)}$ which embeds with distortion D into an ultrametric. Here $\theta = \theta(D) \in (0, 1)$ is the unique solution of the equation*

$$\frac{2}{D} = (1-\theta)\theta^{\frac{\theta}{1-\theta}}.$$

It is elementary to check that $\theta(D) \geqslant 1 - \frac{2e}{D}$ for all $D > 2$, and that as $\varepsilon \searrow 0$ we have $\theta(2+\varepsilon) = \frac{\varepsilon}{2\log(1/\varepsilon)} + O\left(\frac{\varepsilon \log\log(1/\varepsilon)}{(\log(1/\varepsilon))^2}\right)$. Theorem 1.2 yields a very short proof (complete details in 3 pages) of the the nonlinear Dvoretzky theorem for all distortions $D > 2$, with the best known bounds on the exponent $\theta(D)$. In a sense that is made precise in Section 1.2, the above value of $\theta(D)$ is optimal for our method.

1.1. **Approximate distance oracles and limitations of Ramsey partitions.** We recall some terminology and results from [12]. Fix $\delta \in (0,1)$ and let (X,d) be an n-point metric space of diameter 1. A sequence $\{\mathscr{P}_k\}_{k=0}^{\infty}$ of partitions of X is called a *partition tree* of rate δ if $\mathscr{P}_0$ is the trivial partition $\{X\}$, for all $k \geqslant 0$ $\mathscr{P}_{k+1}$ is a refinement of $\mathscr{P}_k$, and each set in $\mathscr{P}_k$ has diameter at most δ^k.

The main tool in [12] is *random partition trees.* Let Pr be a probability distribution over partition trees of rate δ. For $\ell > 0$ consider the the random subset $Y \subseteq X$ consisting of those $x \in X$ such that for *all* $k \in \mathbb{N}$ the entire closed ball $B(x, \delta^k/\ell)$ is contained in the element of $\mathscr{P}_k$ to which x belongs. Assume that each $x \in X$ falls in Y with Pr-probability at least $n^{-\beta}$. Then $\mathbb{E}\left[|Y|\right] \geqslant n^{1-\beta}$. Define for distinct $x, y \in X$ the random quantity $\rho(x,y) = \delta^{k(x,y)}$, where $k(x,y)$ is the largest integer k such that both x and y fall in the same element of $\mathscr{P}_k$. Then ρ is an ultrametric on X, and for $x \in X$ and $y \in Y$ we have $\rho(x,y) \geqslant d(x,y) \geqslant \frac{\delta}{\ell}\rho(x,y)$ [12, Lem. 2.1]. Thus, on Y, ρ is bi-Lipschitz equivalent to the original metric d with distortion $\leqslant \ell/\delta$. But more is true: the ultrametric ρ is defined on all of X, and approximates up to a factor $\leqslant \ell/\delta$ all distances from points of Y to *all the other points of* X.

In [12] random partition trees were constructed with the desired bounds on β and the distortion ℓ/δ. It was shown in [12] that the existence of an ultrametric ρ on X which has the above property of approximating distances from points of a large subset $Y \subseteq X$ to all other points of X, has a variety of implications to the theory of data structures. Here we need to briefly recall the connection to *approximate distance oracles.*

An n-point metric space (X,d) can be thought of as table of $\binom{n}{2}$ numbers, corresponding to the distances between all unordered pairs $x, y \in X$. In the approximate distance oracle problem the goal is, given $D > 1$, to do "one time work" (preprocessing) that produces a data structure (called an approximate distance oracle) of size $o(n^2)$ such that given a "query" $x, y \in X$, one can quickly produce a number $E(x,y)$ satisfying $d(x,y) \leqslant E(x,y) \leqslant Dd(x,y)$. We call D the *stretch* of the approximate distance oracle.

The seminal work on approximate distance oracles is due to Thorup and Zwick [16], who showed that for all odd $D \in \mathbb{N}$ one can design a data structure of size $O(Dn^{1+2/D})$ using which one can compute in time $O(D)$ a number $E(x,y)$ satisfying $d(x,y) \leqslant E(x,y) \leqslant Dd(x,y)$. In [12] it was shown[1] that if every n-point metric space (X,d) admits an ultrametric ρ (defined on all of X) and a subset $Y \subseteq X$ with $|Y| \geqslant n^{1-c/D}$, such that for every $x \in X$ and $y \in Y$ we have $d(x,y) \leqslant \rho(x,y) \leqslant Dd(x,y)$, then any n-point metric space can be preprocessed to yield a data structure of size $O(n^{1+c/D})$ using which one can compute in

[1]This assertion is not stated explicitly in [12], but it follows directly from the proof of [12, Th. 1.2]: using the notation of [12], as noted in the proof of [12, Th. 1.2], the ultrametric ρ_j is only required to be defined, and satisfy the conclusion of [12, Lem. 4.2], on X_{j-1} and not on all of X. This property is guaranteed by our assumption. Thus there is no loss of constant factor since for the purpose of [12, Th. 1.2] (unlike other applications of [12, Lem. 4.2] in [12]), we do not need to use [12, Lem. 4.1].

time $O(1)$ a number $E(x,y)$ satisfying $d(x,y) \leqslant E(x,y) \leqslant Dd(x,y)$. A key new point here is that the query time is a universal constant, and does not depend on D as in [16].

It was also shown in [16] that any approximate distance oracle that answers distance queries with stretch $D < 3$ must use $\gtrsim n^2$ bits of storage. Combining this lower bound with the above construction of [12], we see that if $D < 3$ there must exist arbitrarily large n-point metric spaces (X_n, d_n) such that if ρ is an ultrametric on X_n and $Y \subseteq X_n$ is such that $d(x,y) \leqslant \rho(x,y) \leqslant Dd(x,y)$ for all $x \in X_n$ and $y \in Y$, then $|Y| \lesssim n^{o(1)}$. It is actually not difficult to unravel the arguments of [16, 12] to give a direct proof of the fact that Ramsey partitions cannot yield the nonlinear Dvoretzky theorem for distortions in $(2,3)$. We will not do so here since it would be a digression from the topic of the present paper; the purpose of the above discussion is only to explain why a method other than Ramsey partitions is required in order to to go all the way down to distortion 2.

1.2. **The fragmentation procedure and admissible exponents.** Having realized that a proof of part (2) of Theorem 1.1 for D arbitrarily close to 2 cannot produce a large $Y \subseteq X$ and an ultrametric ρ that is defined on all of X and satisfies $d(x,y) \leqslant \rho(x,y) \leqslant Dd(x,y)$ for all $x \in X$ and $y \in Y$, it is natural to try to design a procedure which results in an ultrametric that is defined on the subset Y alone. This is what our fragmentation procedure does.

In order to state our main results, we require the following definition:

Definition 1.3 (Admissible exponent). *Fix $D > 2$. We say that $\sigma > 0$ is an* admissible exponent *for D if there exist a sequence of (not necessarily independent) random variables $1 = r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$ with $\lim_{n\to\infty} r_n = 0$, such that for every real number $r > 0$, we have*

$$\sum_{n=1}^{\infty} \Pr\left[r_n < r \leqslant r_n + \frac{2r_{n-1}}{D}\right] \leqslant \sigma. \tag{1}$$

Theorem 1.4 (Ultrametrics via admissible exponents). *Fix $D > 2$, and let $\sigma > 0$ be an admissible exponent for D. Let $X = (X,d)$ be a finite metric space. Then there exists a subset S of X of cardinality $|S| \geqslant |X|^{1-\sigma}$ which embeds with distortion D into an ultrametric.*

Let $\sigma^*(D)$ denote the infimum of those $\sigma > 0$ which are admissible exponents for D. Due to Theorem 1.4, we would like to estimate $\sigma^*(D)$. In fact, it turns out that we can compute it exactly; the following theorem, in combination with Theorem 1.4, implies Theorem 1.2.

Theorem 1.5 (Optimization of admissible exponents). *For every $D > 2$ we have $\sigma^*(D) = \beta$, where $\beta \in (0,1)$ is the unique solution of the equation*

$$\frac{2}{D} = \beta(1-\beta)^{\frac{1-\beta}{\beta}}. \tag{2}$$

Moreover, $\sigma^*(D)$ is attained at the following random variables: $r_0 = 1$, and for $n \in \mathbb{N}$,

$$r_n \stackrel{\mathrm{def}}{=} (1-\beta)^{\frac{U+n-1}{\beta}}, \tag{3}$$

where U is a random variable that is uniformly distributed on the interval $[0,1]$. For this choice of $1 = r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$, the supremum of the left hand side of (1) over $r > 0$ equals the value of β in (2).

The construction of the subset S in Theorem 1.4 is most natural to describe in the context of compact metric spaces, though it will be applied here only to finite metric spaces.

Throughout this paper a *metric probability space* (X, d, μ) is a compact metric space (X, d) equipped with a Borel probability measure μ. For $x \in X$ and $r \geqslant 0$ we shall use the standard notation for (closed) balls: $B(x, r) = \{y \in X : d(x, y) \leqslant r\}$. To avoid degeneracies we assume that for every $r > 0$ we have $\mu(B(x, r)) > 0$, and that the function $x \mapsto \mu(B(x, r))$ is measurable. Of course, these hypotheses are automatic in the case of finite metric spaces with uniform measure.

Fix a metric probability space (X, d, μ), normalized to have diameter 2, and a decreasing sequence of radii $1 = r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$. Define inductively a decreasing sequence of random subsets $X = S_0 \supseteq S_1 \supseteq S_2 \ldots$ as follows. Having defined S_i, let $\{x_n\}_{n=1}^{\infty}$ be an i.i.d. sequence of points in X, each distributed according to μ. The set S_{i+1} is defined to be those points $x \in S_i$ for which the first point in the sequence $\{x_n\}_{n=1}^{\infty}$ that fell in $B\left(x, r_n + \frac{2r_{n-1}}{D}\right)$, actually fell in the smaller ball $B(x, r_n)$. Letting $S = \bigcup_{i=0}^{\infty} S_i$, we argue (see Lemma 2.1 and Lemma 2.2) that (S, d) embeds with distortion D into an ultrametric, and that,

$$\mathbb{E}\left[\mu(S)\right] \geqslant \int_X \left(\prod_{n=1}^{\infty} \frac{\mu(B(x, r_n))}{\mu\left(B\left(x, r_n + \frac{2r_{n-1}}{D}\right)\right)} \right) d\mu(x). \tag{4}$$

So far we did not use the fact that the radii $\{r_n\}_{n=0}^{\infty}$ are themselves random. The additional randomness allows us to use a refinement of an idea of [12] in order to control the infinite product appearing in (4) using Jensen's inequality (the corresponding step in [12] used the AM-GM inequality). This is how the notion of admissible exponent appears in Theorem 1.4; the details appear in Section 2. Note that the proof of Theorem 1.2 is simple to describe: it follows the above outline with the specific sequence of random radii given in (3). (Observe that this sequence of radii involves a choice of only one random number U, unlike the construction of [12], and its predecessors [7, 10], in which r_n was uniformly distributed on $[8^{-n}/4, 8^{-n}/2]$, and the $\{r_n\}_{n=0}^{\infty}$ were independent random variables.)

The obvious weakness of the above approach is that the random radii $\{r_n\}_{n=0}^{\infty}$ are chosen without consideration of the particular geometry of the metric space X. It makes sense that in order to obtain sharper results one would need to investigate how different scales in X interact, and reflect this understanding in a choice of radii which are not "scale-oblivious". Theorem 1.5 shows that in order to improve our bounds in Theorem 1.2 one would need to use a fragmentation procedure that is not scale-oblivious (or, find a way to control an expression such as (4) without using Jensen's inequality; this seems quite difficult).

A particular question of interest in this context is as follows: for $D > 2$ let $\theta^*(D)$ be the supremum of those $\theta > 0$ such that there exists $n_0 \in \mathbb{N}$ for which any metric space of cardinality $n \geqslant n_0$ has a subset of size $\geqslant n^{\theta}$ that embeds with distortion D into an ultrametric. Both [3] and our new proof give the bound $\theta^*(2+\varepsilon) \gtrsim \varepsilon/\log(2/\varepsilon)$ (for different reasons). Must it be the case that $\theta^*(2+\varepsilon)$ tends to 0 as $\varepsilon \searrow 0$? This is of course related the unknown behavior of the nonlinear Dvoretzky problem at distortion $D = 2$. Computing the value of $\limsup_{D\to\infty} D(1 - \theta^*(D))$ is also of interest; due to Theorem 1.5 we know that using our scale-oblivious metric fragmentation procedure we cannot bound this number by less than $2e$.

Acknowledgements. We thank Manor Mendel for helpful discussions on the Thorup-Zwick lower bound. A. N. is supported by NSF grants CCF-0635078 and CCF-0832795, BSF grant 2006009, and the Packard Foundation. T. T. is supported by a grant from the MacArthur foundation, by NSF grant DMS-0649473, and by the NSF Waterman award.

2. Randomized fragmentation

We begin with a lemma that fragments a metric space at a single pair of scales $R > r > 0$.

Lemma 2.1 (Fragmentation lemma)**.** *Let (X, d, μ) be a metric probability space, and let $S \subseteq X$ be a compact subset of X. Fix $R > r > 0$ and a Borel-measurable non-negative function $w : S \to [0, \infty)$. Then there exists a compact subset $T \subseteq S$ with*

$$\int_T \frac{\mu(B(x,R))}{\mu(B(x,r))} w(x)\ d\mu(x) \geqslant \int_S w(x)\ d\mu(x), \tag{5}$$

such that T can be partitioned as $T = \bigcup_{n=1}^{\infty} T_n$, where each (possibly empty) T_n is compact and contained in a ball of radius r, and any two non-empty T_n, T_m are separated by a distance of at least $R - r$.

Proof. We use the probabilistic method. Let $\{x_n\}_{n=1}^{\infty}$ be an i.i.d. sequence of points in X, selected using the measure μ. Observe that as $B(x, R)$ has positive measure for all $x \in X$, we will almost surely have $x_n \in B(x, R)$ for at least one $n \in \mathbb{N}$. Thus if we define the (random) quantity

$$n(x) \stackrel{\text{def}}{=} \inf\{n \in \mathbb{N} : x_n \in B(x, R)\}, \tag{6}$$

then $n(x)$ is finite for almost every $x \in X$, and $x \mapsto n(x)$ is a measurable function of x.

Define a (random) subset $A \subseteq S$ by

$$A \stackrel{\text{def}}{=} \{x \in S : n(x) < \infty\ \wedge\ x_{n(x)} \in B(x, r)\}. \tag{7}$$

Then $A = \bigcup_{n=1}^{\infty} A_n$, where

$$A_n \stackrel{\text{def}}{=} \{x \in S : n(x) = n\ \wedge\ x_n \in B(x, r)\}. \tag{8}$$

By definition we have $A_n \subseteq B(x_n, r)$. Also, if $x \in A_n$ and $y \in A_m$ for some $1 \leqslant n < m$, then by the definitions (6), (8) we have $d(x_n, x) \leqslant r$ and $d(x_n, y) > R$, and hence by the triangle inequality we have $d(x, y) > R - r$. Thus if we set $T_n \stackrel{\text{def}}{=} \overline{A_n}$, then T_n and T_m are compact and separated by a distance of at least $R - r$ (this shows that only finitely many of the T_n are non-empty). If we define $T \stackrel{\text{def}}{=} \bigcup_{n=1}^{\infty} T_n$, then T is a compact subset of S.

Since $T \supseteq A$, in order to conclude the proof of Lemma 2.1 it suffices to prove the identity

$$\mathbb{E}\left[\int_A \frac{\mu(B(x,R))}{\mu(B(x,r))} w(x)\ d\mu(x)\right] = \int_S w(x)\ d\mu(x). \tag{9}$$

By the Fubini-Tonelli theorem, in order to prove (9) it suffices to show that for all $x \in S$ we have,

$$\Pr[x \in A] = \frac{\mu(B(x,r))}{\mu(B(x,R))}. \tag{10}$$

Since $n(x)$ is finite almost surely, the definition (7), together with the joint independence of $x_1, x_2 \ldots$, immediately implies that:

$$\Pr[x \in A] = \sum_{n=1}^{\infty} \Pr\left[x_n \in B(x,r)\ \wedge\ x_1, \ldots, x_{n-1} \notin B(x,R)\right]$$
$$= \sum_{n=1}^{\infty} \mu(B(x,r))\,(1 - \mu(B(x,R)))^{n-1} = \frac{\mu(B(x,r))}{\mu(B(x,R))}. \tag{11}$$

This proves (10), and thus concludes the proof of Lemma 2.1. □

We can iterate Lemma 2.1 as follows.

Lemma 2.2 (Iterated fragmentation lemma)**.** *Fix $R, D > 0$. Let (X, d, μ) be a metric probability space of diameter at most $2R$, and let*

$$R = r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$$

be a sequence of radii converging to zero. Then there exists a compact subset S of X such that

$$\mu(S) \geqslant \int_X \left(\prod_{n=1}^{\infty} \frac{\mu(B(x, r_n))}{\mu\left(B\left(x, r_n + \frac{2r_{n-1}}{D}\right)\right)} \right) d\mu(x), \tag{12}$$

and (S, d) embeds with distortion D into an ultrametric.

Proof. By applying Lemma 2.1 repeatedly, we obtain a decreasing sequence of compact subsets of X,

$$X = S_0 \supseteq S_1 \supseteq S_2 \supseteq \ldots$$

satisfying for $n \geqslant 1$,

$$\int_{S_n} \left(\prod_{m=n+1}^{\infty} \frac{\mu(B(x, r_m))}{\mu\left(B\left(x, r_m + \frac{2r_{m-1}}{D}\right)\right)} \right) d\mu(x) \geqslant \int_{S_{n-1}} \left(\prod_{m=n}^{\infty} \frac{\mu(B(x, r_m))}{\mu\left(B\left(x, r_m + \frac{2r_{m-1}}{D}\right)\right)} \right) d\mu(x),$$

such that for each $n \in \mathbb{N}$ we have $S_n = \bigcup_{j=1}^{\infty} S_{n,j}$, where each $S_{n,j}$ is compact and contained in a ball of radius r_n, and if $S_{n,j}, S_{n,\ell} \neq \emptyset$ then $d(S_{n,j}, S_{n,\ell}) \geqslant 2r_{n-1}/D$. It follows inductively that

$$\int_{S_n} \left(\prod_{m=n+1}^{\infty} \frac{\mu(B(x, r_m))}{\mu\left(B\left(x, r_m + \frac{2r_{m-1}}{D}\right)\right)} \right) d\mu(x) \geqslant \int_{X} \left(\prod_{m=1}^{\infty} \frac{\mu(B(x, r_m))}{\mu\left(B\left(x, r_m + \frac{2r_{m-1}}{D}\right)\right)} \right) d\mu(x),$$

and in particular

$$\mu(S_n) \geqslant \int_{X} \left(\prod_{m=1}^{\infty} \frac{\mu(B(x, r_m))}{\mu\left(B\left(x, r_m + \frac{2r_{m-1}}{D}\right)\right)} \right) d\mu(x).$$

If we set $S \stackrel{\text{def}}{=} \bigcap_{n=1}^{\infty} S_n$, then S is compact and obeys (12).

If $x, y \in S$ are distinct, let $n(x, y)$ be the largest integer n such that for all $m \in \{1, \ldots, n\}$ there is $j(m) \in \mathbb{N}$ for which $x, y \in S_{m,j(m)}$. Note that since the diameter of $S_{m,j}$ is at most $2r_m$, and $\lim_{m\to\infty} r_m = 0$, such an n must exist. Now define an ultrametric ρ on S by

$$\rho(x, y) \stackrel{\text{def}}{=} 2r_{n(x,y)}.$$

It is immediate to check that ρ is symmetric, and obeys the ultratriangle inequality

$$\forall x, y, z \in S, \quad \rho(x, z) \leqslant \max\{\rho(x, z), \rho(y, z)\}.$$

If $x, y \in S$ are distinct and $n = n(x, y)$, then by definition $x, y \in S_{n,j}$ for some $j \in \mathbb{N}$ and $x \in S_{n+1,k}$, $y \in S_{n+1,\ell}$, where $k \neq \ell$. Thus $d(x, y) \leqslant \operatorname{diam}(S_{n,j}) \leqslant 2r_n = \rho(x, y)$ and $d(x, y) \geqslant d(S_{n+1,k}, S_{n+1,\ell}) \geqslant 2r_n/D = \rho(x, y)/D$. It follows that the identity map from (S, d) to (S, ρ) has distortion at most D, completing the proof of Lemma 2.2. □

Now suppose that (X, d) is a finite metric space, and that μ is the counting measure on X. Then Lemma 2.2 specializes to

Corollary 2.3 (Iterated fragmentation lemma, finite case)**.** *Fix $D, R > 0$. Let $X = (X, d)$ be a finite metric space of diameter at most $2R$, and let*

$$R = r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$$

be a sequence of radii converging to zero. Then there exists a subset S of X such that

$$|S| \geqslant \sum_{x \in X} \prod_{n=1}^{\infty} \frac{|B(x, r_n)|}{\left|B\left(x, r_n + \frac{2r_{n-1}}{D}\right)\right|}, \tag{13}$$

and (S, d) embeds with distortion D into an ultrametric.

The condition (13) is difficult to work with. However, using a random choice of r_n, and Jensen's inequality, one can obtain a more workable condition in terms of the notion of admissible exponent as in Definition 1.3. This is contained in Theorem 1.4, which we are now in position to prove.

Proof of Theorem 1.4. By rescaling we may assume that X has diameter at most 2. We let $r_0, r_1, \ldots$ be the random variables in Definition 1.3, i.e., (1) holds for all $r > 0$. Applying Corollary 2.3 we thus obtain a (random) subset $S \subseteq X$ obeying (13), which embeds with distortion D into an ultrametric. Taking expectations we obtain

$$\mathbb{E}\left[|S|\right] \geqslant \sum_{x \in X} \mathbb{E}\left[\prod_{n=1}^{\infty} \frac{|B(x, r_n)|}{\left|B\left(x, r_n + \frac{2r_{n-1}}{D}\right)\right|}\right],$$

and hence by Jensen's inequality,

$$\mathbb{E}\left[|S|\right] \geqslant \sum_{x \in X} \exp\left(\mathbb{E}\left[\sum_{n=1}^{\infty} \log\left(\frac{|B(x, r_n)|}{\left|B\left(x, r_n + \frac{2r_{n-1}}{D}\right)\right|}\right)\right]\right). \tag{14}$$

For every $x \in X$ let $0 = t_1(x) < t_2(x) < \ldots < t_{k(x)}(x)$ be the radii at which $|B(x,t)|$ jumps, i.e., $1 = |B(x, t_1(x))| < |B(x, t_2(x))| < \ldots < |B(x, t_{k(x)}(x))| = |X|$, and $B(x,t) = B(x, t_j(x))$ if $t_j(x) \leqslant t < t_{j+1}(x)$ (where we use the convention $t_{k(x)+1}(x) = \infty$). Note that for every random variable $r \geqslant 0$ we have the following simple identity:

$$\begin{aligned}
\mathbb{E}\left[\log |B(x,r)|\right] &= \sum_{j=1}^{k(x)} \Pr\left[t_j(x) \leqslant r < t_{j+1}(x)\right] \log |B(x, t_j(x))| \\
&= \sum_{j=1}^{k(x)} \left(\Pr\left[r \geqslant t_j(x)\right] - \Pr\left[r \geqslant t_{j+1}(x)\right]\right) \log |B(x, t_j(x))| \\
&= \sum_{j=2}^{k(x)} \Pr\left[r \geqslant t_j(x)\right] \log\left(\frac{|B(x, t_j(x))|}{|B(x, t_{j-1}(x))|}\right). \qquad (15)
\end{aligned}$$

Applying (15) to $r = r_n$ and $r = r_n + \frac{2r_{n-1}}{D}$, we see that (14) can be written as

$$\mathbb{E}\left[|S|\right]$$
$$\geqslant \sum_{x \in X} \exp\left(-\sum_{j=2}^{k(x)} \left(\sum_{n=1}^{\infty} \Pr\left[r_n < t_j(x) \leqslant r_n + \frac{2r_{n-1}}{D}\right]\right) \log\left(\frac{|B(x, t_j(x))|}{|B(x, t_{j-1}(x))|}\right)\right). \tag{16}$$

Applying (1) we conclude that

$$\mathbb{E}\left[|S|\right] \geqslant \sum_{x\in X} \exp\left(-\sigma \sum_{j=2}^{k(x)} \log\left(\frac{|B(x,t_j(x))|}{|B(x,t_{j-1}(x))|}\right)\right) = \sum_{x\in X} |X|^{-\sigma} = |X|^{1-\sigma},$$

where we used the fact that $|B(x,t_1(x))| = 1$ and $|B(x,t_{k(x)}(x))| = |X|$. The proof of Theorem 1.4 is complete. □

3. Proof of Theorem 1.5

Define $f:[0,1]\to[0,1]$ by $f(\beta) = \beta(1-\beta)^{\frac{1-\beta}{\beta}}$, where $f(0)=0$ and $f(1)=1$. Note that $(\log f)'(\beta) = -\frac{1}{\beta^2}\log(1-\beta)$, and therefore f is strictly increasing on $[0,1]$. It follows that for each $\alpha\in[0,1]$ there is a unique $\beta=\beta(\alpha)$ satisfying the identity

$$\alpha = \beta(1-\beta)^{\frac{1-\beta}{\beta}}. \tag{17}$$

Fix $D>2$ and set $\beta = \beta(2/D)$. Let U be a random variable that is uniformly distributed on $[0,1]$. We shall define a sequence of random variables $r_0 \geqslant r_1 \geqslant r_2 \geqslant \ldots > 0$ as in (3), i.e., by setting $r_0 = 1$, and for $n\in\mathbb{N}$,

$$r_n \stackrel{\mathrm{def}}{=} (1-\beta)^{\frac{U+n-1}{\beta}}.$$

Writing $\alpha = 2/D$, for every $r>0$ we have the following bound on the left hand side of (1):

$$\begin{aligned}&\sum_{n=1}^{\infty}\Pr\left[r_n < r \leqslant r_n + \frac{2r_{n-1}}{D}\right]\\ &\qquad\leqslant \sum_{n=1}^{\infty}\Pr\left[(1-\beta)^{\frac{U+n-1}{\beta}} < r \leqslant (1-\beta)^{\frac{U+n-1}{\beta}} + \alpha(1-\beta)^{\frac{U+n-2}{\beta}}\right] = \sum_{n=1}^{\infty}\Pr\left[U\in I_n\right], \end{aligned}\tag{18}$$

where I_n is the interval:

$$I_n \stackrel{\mathrm{def}}{=} \left(\frac{\beta\log r}{\log(1-\beta)} - n + 1, \frac{\beta\log r}{\log(1-\beta)} - n + 1 - \frac{\beta\log\left(1+\frac{\alpha}{(1-\beta)^{1/\beta}}\right)}{\log(1-\beta)}\right] \stackrel{\mathrm{def}}{=} (a-n, b-n].$$

The identity (17) implies that $b-a=\beta$. In particular, $b-a\leqslant 1$, and hence the intervals $\{I_n\}_{n=1}^\infty$ are disjoint, at most two of them intersect $[0,1]$, and the total length of the intersection of $\bigcup_{n=1}^\infty I_n$ with $[0,1]$ is at most $b-a$. Combined with (18), this observation implies that

$$\sum_{n=1}^{\infty}\Pr\left[r_n < r \leqslant r_n + \frac{2r_{n-1}}{D}\right] \leqslant \mathrm{length}\left(\left(\bigcup_{n=1}^{\infty} I_n\right)\cap[0,1]\right) \leqslant b-a = \beta.$$

This proves the second assertion of Theorem 1.5. It remains to prove that for all $D>2$ we have $\sigma^*(D)\geqslant\beta(2/D)$. To this end let $\{r_n\}_{n=0}^\infty$ be a sequence of random variables decreasing to zero as in Definition 1.3, so that (1) holds for some $\sigma>0$. Our goal is to show that $\sigma\geqslant\beta(2/D)$.

For $\alpha,p\in(0,1)$ denote

$$\beta_p(\alpha) \stackrel{\mathrm{def}}{=} \inf_{x>1}\frac{(1+\alpha x)^p - 1}{x^p - 1}. \tag{19}$$

By homogeneity, for every $y \geqslant x > 0$ we have $(x+\alpha y)^p - x^p \geqslant \beta_p(\alpha)\,(y^p - x^p)$. Thus for all $n \in \mathbb{N}$,

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\left(r_n + \frac{2r_{n-1}}{D}\right)^p - r_n^p\right] \geqslant \beta_p\left(\frac{2}{D}\right)\mathbb{E}\left[\sum_{n=1}^{\infty}\left(r_{n-1}^p - r_n^p\right)\right] = \beta_p\left(\frac{2}{D}\right), \tag{20}$$

where we used the fact that $\lim_{n\to\infty} r_n = 0$ and $r_0 = 1$.

Now,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{E}\left[\left(r_n + \frac{2r_{n-1}}{D}\right)^p - r_n^p\right] &= \sum_{n=1}^{\infty}\int_0^{\infty} pr^{p-1}\left(\Pr\left[r_n + \frac{2r_{n-1}}{D} \geqslant r\right] - \Pr\left[r_n \geqslant r\right]\right) dr \\ &= \sum_{n=1}^{\infty}\int_0^{1+\frac{2}{D}} pr^{p-1}\Pr\left[r_n < r \leqslant r_n + \frac{2r_{n-1}}{D}\right] dr \overset{(1)}{\leqslant} \sigma\left(1+\frac{2}{D}\right)^p. \end{aligned}\tag{21}$$

By combining (20) and (21) (which hold for all $p \in (0,1)$), we see that the bound $\sigma \geqslant \beta(2/D)$ will be proven if we manage to show that for all $\alpha \in (0,1)$,

$$\limsup_{p\to 0} \beta_p(\alpha) \geqslant \beta(\alpha), \tag{22}$$

where $\beta(\alpha)$ is the unique $\beta \in (0,1)$ satisfying (17).

To prove (22), define $f : [0,\infty) \to \mathbb{R}$ by $f(x) = (1+\alpha x)^p - 1 - \beta_p(\alpha)\,(x^p - 1)$. Note that f takes only non-negative values, due to the definition (19). By considering the limit as $x \to \infty$ of the right hand side of (19), we see that $\beta_p(\alpha) \leqslant \alpha^p$. But, it cannot be the case that $\beta_p(\alpha) = \alpha^p$, since otherwise $f(x) = (1+\alpha x)^p - (\alpha x)^p - (1-\alpha^p)$, which, since $p \in (0,1)$, tends to $-(1-\alpha^p) < 0$ as $x \to \infty$, contradicting the non-negativity of f on $[1,\infty)$. Thus $\beta_p(\alpha) < \alpha^p$. It follows in particular that the infimum in (19) is actually a minimum, i.e., there exists $x_0 \in (1,\infty)$ for which $\beta_p(\alpha) = \frac{(1+\alpha x_0)^p - 1}{x_0^p - 1}$. This is the same as $f(x_0) = 0$, and since f is non-negative, x_0 must be a global minimum of f, and hence $f'(x_0) = 0$.

From $0 = f'(x_0) = p\alpha(1+\alpha x_0)^{p-1} - p\beta_p(\alpha)x_0^{p-1}$ we see that

$$x_0 = \frac{1}{\left(\alpha/\beta_p(\alpha)\right)^{1/(1-p)} - \alpha}. \tag{23}$$

Substituting this value of x_0 into the equation $f(x_0) = 0$, we see that

$$\frac{\left(\alpha/\beta_p(\alpha)\right)^{p/(1-p)}}{\left(\left(\alpha/\beta_p(\alpha)\right)^{1/(1-p)} - \alpha\right)^p} - 1 = \beta_p(\alpha)\left(\frac{1}{\left(\left(\alpha/\beta_p(\alpha)\right)^{1/(1-p)} - \alpha\right)^p} - 1\right). \tag{24}$$

Denote $\beta = \limsup_{p\to 0}\beta_p(\alpha)$. If $\beta = 1$ then (22) holds trivially. We may therefore assume that $\beta < 1$. Moreover, (23) combined with $x_0 > 1$ implies that $\left(\alpha/\beta_p(\alpha)\right)^{1/(1-p)} - \alpha < 1$, or $\beta_p(\alpha) > \frac{\alpha}{(1+\alpha)^{1-p}}$. Thus $\beta \geqslant \frac{\alpha}{1+\alpha}$ (all that we will need below is that $\beta \neq 0$).

If $\{p_k\}_{k=1}^{\infty} \subseteq (0,1)$ is such that $\lim_{k\to\infty} p_k = 0$ and $\lim_{k\to\infty}\beta_{p_k}(\alpha) = \beta$, then it follows from (24) that:

$$p_k\left(\log\left(\frac{\alpha}{\beta}\right) - \log\left(\frac{\alpha}{\beta} - \alpha\right)\right) + o(p_k) = -p_k\beta\log\left(\frac{\alpha}{\beta} - \alpha\right) + o(p_k). \tag{25}$$

Since, as argued above, $\frac{\alpha}{\beta}, \frac{\alpha}{\beta} - \alpha \in (0, \infty)$, the asymptotic identity (25) implies that

$$\log\left(\frac{\alpha}{\beta}\right) - \log\left(\frac{\alpha}{\beta} - \alpha\right) = -\beta \log\left(\frac{\alpha}{\beta} - \alpha\right),$$

which simplifies to give $\alpha = \beta(1-\beta)^{(1-\beta)/\beta}$. Since we already argued (in the paragraph preceding (17)), that $\beta(\alpha)$ is the unique solution of the equation (17), we deduce that $\beta = \beta(\alpha)$. The proof of (22), and hence also the proof of Theorem 1.5, is complete. $\square$

References

[1] Y. Bartal, B. Bollobás, and M. Mendel. Ramsey-type theorems for metric spaces with applications to online problems. *J. Comput. System Sci.*, 72(5):890–921, 2006.

[2] Y. Bartal, N. Linial, M. Mendel, and A. Naor. On metric Ramsey-type dichotomies. *J. London Math. Soc. (2)*, 71(2):289–303, 2005.

[3] Y. Bartal, N. Linial, M. Mendel, and A. Naor. On metric Ramsey-type phenomena. *Ann. of Math. (2)*, 162(2):643–709, 2005.

[4] Y. Bartal, N. Linial, M. Mendel, and A. Naor. Some low distortion metric Ramsey problems. *Discrete Comput. Geom.*, 33(1):27–41, 2005.

[5] A. Blum, H. Karloff, Y. Rabani, and M. Saks. A decomposition theorem for task systems and bounds for randomized server problems. *SIAM J. Comput.*, 30(5):1624–1661 (electronic), 2000.

[6] J. Bourgain, T. Figiel, and V. Milman. On Hilbertian subsets of finite metric spaces. *Israel J. Math.*, 55(2):147–152, 1986.

[7] G. Calinescu, H. Karloff, and Y. Rabani. Approximation algorithms for the 0-extension problem. *SIAM J. Comput.*, 34(2):358–372 (electronic), 2004/05.

[8] M. Charikar and A. Karagiozova. A tight threshold for metric Ramsey phenomena. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 129–136 (electronic), New York, 2005. ACM.

[9] A. Dvoretzky. Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)*, pages 123–160. Jerusalem Academic Press, Jerusalem, 1961.

[10] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *J. Comput. System Sci.*, 69(3):485–497, 2004.

[11] H. Karloff, Y. Rabani, and Y. Ravid. Lower bounds for randomized k-server and motion-planning algorithms. *SIAM J. Comput.*, 23(2):293–312, 1994.

[12] M. Mendel and A. Naor. Ramsey partitions and proximity data structures. *J. Eur. Math. Soc.*, 9(2):253–275, 2007.

[13] V. Milman and G. Schechtman. An "isomorphic" version of Dvoretzky's theorem. II. In *Convex geometric analysis (Berkeley, CA, 1996)*, volume 34 of *Math. Sci. Res. Inst. Publ.*, pages 159–164. Cambridge Univ. Press, Cambridge, 1999.

[14] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

[15] G. Schechtman. Two observations regarding embedding subsets of Euclidean spaces in normed spaces. *Adv. Math.*, 200(1):125–135, 2006.

[16] M. Thorup and U. Zwick. Approximate distance oracles. *J. ACM*, 52(1):1–24 (electronic), 2005.

[17] I. A. Vestfrid and A. F. Timan. A universality property of Hilbert spaces. *Dokl. Akad. Nauk SSSR*, 246(3):528–530, 1979.

Courant Institute, New York University, New York NY 10012
E-mail address: `naor@cims.nyu.edu`

Department of Mathematics, UCLA, Los Angeles CA 90095-1555
E-mail address: `tao@math.ucla.edu`