Lecture 10: Bourgain's Theorem

Lecturer: *Sanjeev Arora* Scribe: *Daniel J. Peng*

The goal of this week's lecture is to prove the ℓ_2 version of Bourgain's Theorem:

THEOREM 1 (BOURGAIN)
Every n-point metric embeds into ℓ_2 with distortion $O(\lg n)$.

The proof will depend on the partitioning idea from Fakcharoenphol, Rao and Talwar (2003). The idea is similar to the proof of Bourgain's Theorem for ℓ_1 from Lecture 3. For each length scale s, we define a partition P_s by the following algorithm:

1. Pick uniformly at random a number $R \in [2^{s-1}, 2^s)$.
2. Pick uniformly at random an order σ on the elements of X.
3. Partition the items of X into at most $n = |X|$ blocks as follows.
   (a) Proceed with the elements of X according to the order σ.
   (b) For each element $x \in X$, pick all non-assigned elements within distance R from it, and form a new block.

   We call P_s the partition created above, and denote by $P_s(x)$ the block in which x was placed.

Just as in the ℓ_1 case, we can apply the Padded Decomposition Property.

THEOREM 2 (PADDED DECOMPOSITION PROPERTY)
Let P_s be a partition of X, and let $x \in X$. If

$$\tau \leq \frac{2^{s-3}}{\lg \frac{|B(x, 2^{s+1})|}{|B(x, 2^{s-3})|}}, \tag{1}$$

then $\mathbf{Pr}_{\sigma,R}[B(x,\tau) \subseteq P_s(x)] \geq \frac{1}{2}$.

PROOF: We proved this theorem in lecture 3. □

We also have a corollary; since the growth ratio is less than the number of nodes, that is,

$$\frac{|B(x, 2^{s+1})|}{|B(x, 2^{s-3})|} \leq n, \tag{2}$$

the corollary follows:

COROLLARY 3 (COROLLARY TO THE PADDED DECOMPOSITION PROPERTY)
Let P_s be a partition of X. For all $x \in X$ and any constant $c \in \mathbf{R}$, we have

$$\mathbf{Pr}_{\sigma,R}\left[B\left(x, \frac{2^s}{|c \lg n|}\right) \subseteq P_s(x)\right] \geq \frac{1}{2}. \tag{3}$$

Before proving Bourgain's Theorem, we first establish a simpler result, embedding an n-point metric into ℓ_2 with distortion $O\left((\lg n)^{\frac{3}{2}}\right)$.

1 Simpler Embedding into ℓ_2 with distortion $O\left((\lg n)^{(3/2)}\right)$.

First, we construct "zero sets," so named because they will have coordinate 0 in the embedding.

DEFINITION 1 (ZERO SETS) *To construct the zero set Z_s, merge nodes whose distance is less than $\frac{2^s}{10n}$.*[1] *Then, pick each block in P_s independently with probability $\frac{1}{2}$, and take the union.*

Now, we proceed to construct the embedding into ℓ_2. First, scale the distance function so that the minimum distance is 1, and let Δ be the maximum distance.

Then, since the radius R of a block for Z_s is at least 2^{s-1}, there are $\lg \Delta$ nontrivial zero sets: $Z_1, \ldots, Z_{\lg \Delta}$. Then, we construct the embedding f by Frechet's technique:

DEFINITION 2 *Let f be the function from X into $\mathbf{R}^{\lg \Delta}$ as follows:*

$$f : x \mapsto (d(x, Z_1), d(x, Z_2), \ldots, d(x, Z_{\lg \Delta})) \tag{4}$$

We claim that this embedding f has $O\left((\lg n)^{(3/2)}\right)$ distortion, that is,

$$d(x, y) \leq \mathbf{E}[|f(x) - f(y)|] \leq (\lg n)^{(3/2)} d(x, y). \tag{5}$$

This is immediate from the following theorem:

THEOREM 4
If (X, d) is an n-point metric space and f is an embedding as described above, then for all $x, y \in X$, we have:

$$\frac{d(x, y)^2}{(\lg n)^2} \leq (\mathbf{E}[|f(x) - f(y)|])^2 \leq (\lg n) d(x, y)^2. \tag{6}$$

PROOF: The trivial upper bound follows from the triangle inequality:

$$|f(x) - f(y)| = \sqrt{\sum_{s=1}^{\lg \Delta} |d(x, Z_s) - d(y, Z_s)|^2} \tag{7}$$

$$\leq \sqrt{\sum_{s=1}^{\lg \Delta} |d(x, y)|^2} \tag{8}$$

$$\leq d(x, y)\sqrt{\lg \Delta} \tag{9}$$

so $(\mathbf{E}[|f(x) - f(y)|])^2 \leq \left(d(x, y)\sqrt{\lg \Delta}\right)^2 \leq d(x, y)^2 (\lg \Delta)$.

We improve this bound by observing that only $\lg n$ terms contribute to the summation in Equation 7. Nodes x and y are merged if $2^s \geq 10n(d(x, y))$, so the contribution is zero. On the

[1] That is, if $d(x, y) < \frac{2^s}{10n}$, then merge them into one node z. The metric on the new set is is the shortest path metric d', where for any u,v, $d'(u, z) = \min(d(u, x), d(u, y))$ and $d'(u, v) = \min(d'(u, z) + d'(z, v), d(u, v))$.

other hand, when $2^s \leq d(x,y)$, the distance $d(x,Z_s)$ and $d(y,Z_s)$ will tend to be less than 2^s, so this contribution falls off geometrically. Thus, the terms matter only for the $\lg n$ values of s satisfying $d(x,y)) \leq 2^s \leq 10n(d(x,y))$. We conclude that $(\mathbf{E}[|f(x)-f(y)|])^2 \leq d(x,y)^2(\lg n)$.

For the lower bound, we show that just one coordinate's contribution makes the required contribution. Fix $x \in X$ and $y \in X$. Consider the coordinate s where $2^s \approx \frac{d(x,y)}{4}$.

Since the diameter of each block in P_s is at most 2^{s+1} and the distance $d(x,y)$ is $4(2^s) = 2^{s+2}$, it must be the case that x and y are in different blocks. Thus, the zero set Z_s contains x with probability $\frac{1}{2}$, and it contains y with independent probability $\frac{1}{2}$. By Corollary 3 to the Padded Decomposition Property, we have that $B(x, \frac{2^s}{10\lg n}) \subseteq P(x)$ with probability at least $\frac{1}{2}$. So, with probability $\frac{1}{8}$, the zero set Z_s contains x but not y, and $d(y,Z_s) \geq \frac{2^s}{10\lg n}$. Since $d(x,y) \approx 2^{s+2}$, we have $d(y,Z_s) \geq \frac{d(x,y)}{40\lg n}$ with probability $\frac{1}{8}$. It follows that

$$(\mathbf{E}[|f(x)-f(y)|])^2 \geq \frac{d(x,y)^2}{8(40\lg n)^2}. \tag{10}$$

□

Remark 1 To show this property without the expectation, we simply repeat the process and concatenate the embeddings from each iteration. Chernoff bounds show that this process will bring the embedding arbitrarily close $(1+\epsilon)$ to the expectation.

Remark 2 The scale where $2^s \approx d(x,y)$ is important because it is the only scale where the distances are large enough to make a real contribution, but small enough that x and y are in different blocks.

2 Bourgain's ℓ_2 Theorem (full)

We will now prove Bourgain's theorem: any n-point metric can be embedded into ℓ_2 with $O(\lg n)$ distortion. This will be immediate from the following theorem:

Theorem 5
If (X,d) is an n-point metric space and f is an embedding as described below, then for all $x,y \in X$, we have:

$$\frac{d(x,y)^2}{\lg n} \leq (\mathbf{E}[|f(x)-f(y)|])^2 \leq (\lg n)d(x,y)^2. \tag{11}$$

To prove this theorem, we first apply a technique due to KLMN 2004; we "glue" the $\lg \Delta$ scales into $\lg n$ coordinates.

Definition 3 *Let $R(x,t)$ be the maximum radius R for which $|B(x,R)| \leq 2^t$.*

Definition 4 *Let $K(x,t) = \lceil \lg R(x,t) \rceil$.*

We now define a growth ratio that reflects how quickly the density of vertices changes around x.

Definition 5 *For any small natural numbers c and c', we let $GR = \lg \frac{\left|B(x,2^{m+c'})\right|}{|B(x,2^{m-c})|}$.*

This definition is ambiguous, but we will fix this later.

Remark 3 Observe that if $\lg |B(x, 2^m)| \gg \lg \left|B(x, 2^{m-3})\right|$, then $R(x,t)$ stays around 2^m for many values of t. More precisely, $R(x,t) \approx 2^m$ (and $K(x,t) \approx m$) for about $\lg \frac{|B(x,2^m)|}{|B(x,2^{m-3})|} = \lg(GR)$ values of t.

We have Z_s as before, except that we do not need to merge nodes before constructing Z_s. The Z_s will not be the zero sets for this theorem, however. Instead, we will now define the zero sets W_t as follows, by "gluing" the Z_s together. In order to decide whether or not to join W_t, each node x "sniffs" around its neighborhood to determine $K(x,t)$ and then checks if it lies in $Z_{K(x,t)}$. If so, it joins W_t.

Definition 6 $W_t = \left\{x : x \in Z_{K(x,t)}\right\}$.

This is not the precise definition we will finally use, but it will convey the general idea of the proof. We will define W_t more precisely later.

Remark 4 Observe that when $t = \lg n$, the maximum radius R for which $|B(x,R)| \leq 2^t$ is the maximum distance Δ. Thus, t goes from 1 to $\lg n$, the function $R(x,t)$ goes to Δ, the function $K(x,t)$ goes to $\lg \Delta$, and $Z_{K(x,t)}$ goes to $Z_{\lg \Delta}$, as we would expect.

Again, we define the embedding function f in Frechet's style:

Definition 7 *Let f be the function from X into $\mathbf{R}^{\lg \Delta}$ as follows:*

$$f : x \mapsto (d(x, W_1), d(x, W_2), \ldots, d(x, W_{\lg n})) \tag{12}$$

We now present the proof of Theorem 5.

Proof: The upper bound follows trivially from the triangle inequality:

$$|f(x) - f(y)|^2 = \sum_{t=1}^{\lg n} |d(x, W_t) - d(y, W_t)|^2 \tag{13}$$

$$\leq \sum_{t=1}^{\lg n} |d(x,y)|^2 \tag{14}$$

$$\leq (\lg n) d(x,y)^2 \tag{15}$$

For the lower bound, we consider the scale m where $2^m \approx d(x,y)$. This is the important scale that we noted in Remark 2; however, in contrast to the $O((\lg n)^{3/2})$ embedding, the "gluing" now gives us multiple coordinates that involve this scale. In the previous embedding, there was just one zero set, Z_m, that involved this scale; now, since Remark 3 gives us $\lg(GR)$ values of t for which $K(x,t) \approx m$, we have $\lg(GR)$ zero sets $W_t = \left\{x : x \in Z_{K(x,t)}\right\}$ that involve $Z_m = Z_{K(x,t)}$.

We apply the Padded Decomposition Property to show that $B(x, \frac{2^m}{\lg GR}) \subseteq P_m(x)$ with probability at least $\frac{1}{2}$. Then, we apply the same logic from the previous proof to each coordinate: with probability $\frac{1}{8}$, the zero set W_t contains x but not y, and $d(y, W_t) \geq \frac{2^s}{\lg(GR)} \approx \frac{d(x,y)}{\lg(GR)}$. Thus, it follows that for each t for which $K(x,t) \approx m$, we have:

$$\mathbf{E}[|d(x, W_t) - d(y, W_t)|] \geq \frac{d(x,y)}{\lg(GR)}. \tag{16}$$

This allows us to conclude the proof of the upper bound:

$$(\mathbf{E}[|f(x)-f(y)|])^2 \geq \mathbf{E}[|f(x)-f(y)|^2] \tag{17}$$

$$\geq \sum_{t=1}^{\lg n} \mathbf{E}[|d(x,W_t)-d(y,W_t)|^2] \tag{18}$$

$$\geq \sum_{t:K(x,t)\approx m} \mathbf{E}[|d(x,W_t)-d(y,W_t)|^2] \tag{19}$$

By the padding property from Equation (16):

$$\geq \sum_{t:K(x,t)\approx m} \frac{d(x,y)^2}{(\lg(GR))^2} \tag{20}$$

Since there are $\log(GR)$ such coordinates:

$$\geq (\lg(GR))\frac{d(x,y)^2}{(\lg(GR))^2} \tag{21}$$

$$\geq \frac{d(x,y)^2}{\lg(GR)} \tag{22}$$

Since the growth ratio is at most n:

$$\geq \frac{d(x,y)^2}{\lg n} \tag{23}$$

This proves the theorem. □

There are two subtleties that we overlooked in the proof of the theorem. We consider them here.

SUBTLETY 1: The Padded Decomposition Property does not strictly apply to W_t. A point x is not in W_t because it is not in Z_m, where $m = K(x,t)$. The point x hopes that, with good probability, it is a distance at least $2^m/\log GR$ from W_t, so that it will be far away from a distant point y, $d(x,y) > 2^{m+1}$ that lands in W_t. This hope is jeopardized by the possibility that a point z in $B(x,\tau)$ might be in W_t because it is deciding "looking" at a different $Z_{m'}$, where $m' = K(z,t)$ and m is not necessarily the same as m'. We deal with this by showing that the $K(\cdot,\cdot)$ has a certain "smoothness" property which implies that m is necessarily close to m'.

LEMMA 6 (SMOOTHNESS LEMMA)
Let $x \in X$ and $m = K(x,t)$. Let $z \in B(x,\tau)$, where $\tau \leq \frac{2^m}{10}$. Finally, let $m' = K(z,t)$. Then

$$m' \in \{m-4, m-3, m-2, m-1, m, m+1\}.$$

PROOF: We know that $B(x,2^m)$ contains roughly 2^t points. Since $d(x,z) \leq 2^m$, we know that $B(x,2^m) \subset B(z,2^m+2^m) = B(z,2^{m+1})$, and there must be 2^t points in $B(z,2^{m+1})$. Thus, $m' \leq m+1$.

For the lower bound, we know that $B(z, 2^{m'})$ contains roughly 2^t points. Since $d(x,z) \le \frac{2^m}{10}$, we know that $B(z, 2^{m'}) \subset B(x, 2^{m'} + \frac{2^m}{10})$. □

So, the concern is warranted, but there are only 6 different scales to worry about. Thus, we simply assert that with probability 2^{-6}, we have

$$d(x, Z_{m'}) \ge \frac{2^{m'}}{\lg \frac{\left|B(x, 2^{m'+1})\right|}{\left|B(x, 2^{m'-3})\right|}} \tag{24}$$

for *all* m' between $m-4$ and $m+1$. This simply reduces the constant factor in our padding for W_t in Equation 16.

SUBTLETY 2: The growth ratio GR stands for all of the different growth ratios $\frac{\left|B(x, 2^{m+c'})\right|}{\left|B(x, 2^{m-c})\right|}$ with different constants c and c'. Most notably, in the size of the paddings (Equation 24), we have various m', and in the number of coordinates (Remark 3), the growth ratios have different constants.

We resolve this by picking a large constant c. Then, we expand the number of coordinates by a factor of $2c+1$ by defining the zero sets:

$$W_{i,t} = x : x \in Z_{K(x,t)-i},$$

where i ranges from $-c$ to c and t ranges from 1 to $\lg n$.

Since $K(x,t) - i \approx m$ when $K(x,t) \in [m-c, m+c]$, the summation in Equation 19 includes

$$\lg \frac{\left|B(x, 2^{m+c})\right|}{\left|B(x, 2^{m-c})\right|}$$

values of (i,t). By making c sufficiently large, we can add more and more coordinates into the sum, until we cancel a $\lg(GR)$ in the denominator in Equation 21:

$$(\mathbf{E}[|f(x) - f(y)|])^2 \ge \sum_{t:K(x,t)-i\approx m} \frac{d(x,y)^2}{(\lg(GR))^2} \tag{25}$$

$$\ge \left(\lg \frac{\left|B(x, 2^{m+c})\right|}{\left|B(x, 2^{m-c})\right|}\right) \frac{d(x,y)^2}{(\lg(GR))^2} \tag{26}$$

$$\ge \frac{d(x,y)^2}{\lg(GR)} \tag{27}$$

$$\ge \frac{d(x,y)^2}{\lg n} \tag{28}$$