# Lecture 10: Proof of Bourgain's Theorem

*In which we prove Bourgain's theorem.*

Today we prove the following theorem.

**Theorem 1 (Bourgain)** *Let $d : V \times V \to \mathbb{R}$ be a semimetric defined over a finite set $V$. Then there exists a mapping $F : V \to \mathbb{R}^m$ such that, for every two elements $u, v \in R$,*

$$||F(u) - F(v)||_1 \leq d(u,v) \leq ||F(u) - F(v)||_1 \cdot c \cdot \log |V|$$

*where $c$ is an absolute constant. Given $d$, the mapping $F$ can be found with high probability in randomized polynomial time in $|V|$.*

Together with the results that we proved in the last lecture, this implies that an optimal solution to the Leighton-Rao relaxation can be rounded to an $O(\log n)$-approximate solution to the sparsest cut problem. This was the best known approximation algorithm for sparsest cut for 15 years, until the Arora-Rao-Vazirani algorithm, which will be our next topic.

The theorem has a rather short proof, but there is an element of "magic" to it. We will discuss several examples and we will see what approaches are suggested by the examples. At the end of the discussion, we will see the final proof as, hopefully, the "natural" outcome of the study of such examples and failed attempts.

## 1 Preliminary and Motivating Examples

A first observation is that embeddings of finite sets of points into L1 can be equivalently characterized as probabilistic embeddings into the real line.

**Fact 2** *For every finite set $V$, dimension $m$, and mapping $F : V \to \mathbb{R}^m$, there is a finitely-supported probability distribution $D$ over functions $f : V \to \mathbb{R}$ such that for every two points $u, v \in V$:*

$$\mathop{\mathbb{E}}_{f\sim D} |f(u) - f(v)| = ||F(u) - F(v)||_1$$

*Conversely, for every finite set $V$ and finitely supported distribution $D$ over functions $f : V \to \mathbb{R}$, there is a dimension $m$ and a mapping $F : V \to \mathbb{R}^m$ such that*

$$\mathop{\mathbb{E}}_{f\sim D} |f(u) - f(v)| = ||F(u) - F(v)||_1$$

PROOF: For the first claim, we write $F_i(v)$ for the $i$-th coordinate of $F(v)$, that is $F(v) = (F_1(v), \ldots, F_m(v))$, and we define $D$ to be the uniform distribution over the $m$ functions of the form $x \to m \cdot F_i(x)$.

For the second claim, if the support of $D$ is the set of functions $\{f_1, \ldots, f_m\}$, where function $f_i$ has probability $p_i$, then we define $F(v) := (p_1 f_1(v), \ldots, p_m f_m(v))$. □

It will be easier to reason about probabilistic mappings into the line, so we will switch to the latter setting from now on.

Our task is to associate a number to every point $v$, and the information that we have about $v$ is the list of distances $\{d(u, v)\}$. Probably the first idea that comes to mind is to pick a random reference vertex $r \in V$, and work with the mapping $v \to d(r, v)$, possibly scaled by a multiplicative constant. (Equivalently, we can think about the deterministic mapping $V \to \mathbb{R}^{|V|}$, in which the vertex $v$ is mapped to the sequence $(d(u_1, v), \ldots, d(u_n, v)$, for some enumeration $u_1, \ldots, u_n$ of the elements of $V$.)

This works in certain simple cases.

**Example 3 (Cycle)** *Suppose that $d(\cdot, \cdot)$ is the shortest-path metric on a cycle, we can see that, for every two points on the cycle, $\mathbb{E}_{r\in V} |d(r, u) - d(r, v)|$ is within a constant factor of their distance $d(u, v)$. (Try proving it rigorously!)*

**Example 4 (Simplex)** *Suppose that $d(u, v) = 1$ for every $u \neq v$, and $d(u, u) = 0$. Then, for every $u \neq v$, we have $\mathbb{E}_{r\in V} |d(r, u) - d(r, v)| = \mathbb{P}[r = u \vee r = v] = 2/n$, so, up to scaling, the mapping incurs no error at all.*

But there are also simple examples in which this works very badly.

**Example 5 (1-2 Metric)** *Suppose that for every $u \neq v$ we have $d(u, v) \in \{1, 2\}$ (any distance function satisfying this property is always a metric) and that, in particular, there is a special vertex $z$ at distance 2 from all other vertices, while all other vertices are at distance 1 from each other. Then, for vertices $u, v$ both different from $z$ we have, as before*

$$\mathbb{E}[|d(r,u) - d(r,v)|] = \frac{2}{n}$$

*but for every $v$ different from $z$ we have*

$$\mathbb{E}[|d(r,z) - d(r,v)|] = \frac{n-2}{n} \cdot |2-1| + \frac{1}{n} \cdot |2-0| + \frac{1}{n} \cdot |0-2| = 1 + \frac{2}{n}$$

*and so our error is going to be $\Omega(n)$ instead of the $O(\log n)$ that we are trying to establish.*

Maybe the next simplest idea is that we should pick at random several reference points $r_1, \ldots, r_k$. But how do we combine the information $d(r_1, u), \ldots, d(r_k, u)$ into a single number to associate to $u$? If we just take the sum of the distances, we are back to the case of sampling a single reference point. (We are just scaling up the expectation by a factor of $k$.)

The next simplest way to combine the information is to take either the maximum or the minimum. If we take the minimum, we see that we have the very nice property that we immediately guarantee that our distances in the L1 embedding are no bigger than the original distances, so that it "only" remains to prove that the distances don't get compressed too much.

**Fact 6** *Let $d : V \times V \to \mathbb{R}$ be a semimetric and $A \subseteq V$ be a non-empty subset of points. Define $f_A : V \to \mathbb{R}$ as*

$$f_A(v) := \min_{r \in A} d(r, v)$$

*Then, for every two points $u, v$ we have*

$$|f_A(u) - f_A(v)| \leq d(u, v)$$

PROOF: Let $a$ be the point such that $d(a, u) = f_A(u)$ and $b$ be the point such that $d(b, v) = f_A(v)$. (It's possible that $a = b$.) Then

$$f_A(u) = d(a,u) \geq d(v,a) - d(u,v) \geq d(v,b) - d(u,v) = f_A(v) - d(u,v)$$

and, similarly,

$$f_A(v) = d(b,v) \geq d(u,b) - d(u,v) \geq d(u,a) - d(u,v) = f_A(u) - d(u,v)$$

□

Is there a way to sample a set $A = \{r_1, \ldots, r_k\}$ such that, for every two points $u, v$, the expectation $\mathbb{E}\,|f_A(u) - f_A(v)|$ is not too much smaller than $d(u, v)$? How large should the set $A$ be?

**Example 7 (1-2 Metric Again)** *Suppose that for every $u \neq v$ we have $d(u,v) \in \{1,2\}$, and that we pick a subset $A \subseteq V$ uniformly at random, that is, each event $r \in A$ has probability $1/2$ and the events are mutually independent.*

*Then for every $u \neq v$:*

$$\frac{1}{4} \cdot d(u,v) \leq |\,\mathbb{E}\,|f_A(u) - f_A(v)| \leq d(u,v)$$

*because with probability $1/2$ the set $A$ contains exactly one of the elements $u, v$, and conditioned on that event we have $|f_A(u) - f_A(v)| \geq 1$ (because one of $f_A(u), f_A(v)$ is zero and the other is at least one), which is at least $d(u,v)/2$.*

If we pick $A$ uniformly at random, however, we incur an $\Omega(n)$ distortion in the case of the shortest path metric on the cycle. In all the examples seen so far, we can achieve constant distortion if we "mix" the distribution in which $A$ is a random set of size 1 and the one in which $A$ is a chosen uniformly at random among all sets, say by sampling from the former probability with probability $1/2$ and from the latter with probability $1/2$.

**Example 8 (Far-Away Clusters)** *Suppose now that $d(\cdot,\cdot)$ has the following structure: $V$ is partitioned into clusters $B_1, \ldots, B_k$, where $|B_i| = i$ (so $k \approx \sqrt{2n}$), and we have $d(u,v) = 1$ for vertices in the same cluster, and $d(u,v) = n$ for vertices in different clusters.*

*If $u, v$ are in the same cluster, then $d(u,v) = 1$ and*

$$\mathbb{E}\,|f_A(u) - f_A(v)| = \mathbb{P}[A \text{ contains exactly one of } u, v]$$

*If $u, v$ are in different clusters $B_i, B_j$, then $d(u,v) = n$ and*

$$\mathbb{E}\,|f_A(u) - f_A(v)| \approx n\,\mathbb{P}[A \text{ intersects exactly one of } B_i, B_j]$$

If we want to stick to this approach of picking a set $A$ of reference elements according to a certain distribution, and then defining the map $f_A(v) := \min_{r \in A} d(r,v)$, then the set $A$ must have the property that for every two sets $S, T$, there is at least a probability $p$ that $A$ intersects exactly one of $S, T$, and we would like $p$ to be as large as possible, because the distortion caused by the mapping will be at least $1/p$.

This suggest the following distribution $D$:

1. Sample $t$ uniformly at random in $\{0, \ldots, \log_2 n\}$
2. Sample $A \subseteq V$ by selecting each $v \in V$, independently, to be in $A$ with probability $2^{-t}$ and to be in $V - A$ with probability $1 - 2^{-t}$.

This distribution guarantees the above property with $p = 1/O(\log n)$.

Indeed, the above distribution guarantees a distortion at most $O(\log n)$ in all the examples encountered so far, including the tricky example of the clusters of different size. In each example, in fact, we can prove the following claim: for every two vertices $u, v$, there is a scale $t$, such that conditioned on that scale being chosen, the expectation of $|f_A(u), f_A(v)|$ is at least a constant times $d(u, v)$. We could try to prove Bourgain's theorem by showing that this is true in every semimetric.

Let us call $D_t$ the conditional distribution of $D$ conditioned on the choice of a scale $t$. We would like to prove that for every semimetric $d(\cdot, \cdot)$ and every two points $u, v$ there is a scale $t$ such that

$$\mathop{\mathbb{E}}_{A \sim D_t} |f_A(u) - f_A(v)| \geq \Omega(d(u, v))$$

which, recalling that $|f_A(u) - f_A(v)| \leq d(u, v)$ for every set $A$, is equivalent to arguing that

$$\mathop{\mathbb{P}}_{A \sim D_t} [|f_A(u) - f_A(v)| \geq \Omega(d(u, v))] \geq \Omega(1)$$

For this to be true, there must be distances $d_1, d_2$ such that $d_1 - d_2 \geq \Omega(d(u, v))$ and such that, with constant probability according to $D_t$, we have $f_A(u) \geq d_1$ and $f_A(v) \leq d_2$ (or vice-versa). For this to happen, there must be a constant probability that $A$ avoids the set $\{r : d(u, r) < d_1\}$ and intersects the set $\{r : d(v, r) \leq d_2\}$. For this to happen, both sets must have size $\approx 2^t$.

This means that if we want to make this "at least one good scale for every pair of points" argument work, we need to show that for every two vertices $u, v$ there is a "large" distance $d_1$ and a "small" distance $d_2$ (whose difference is a constant times $d(u, v)$) such that a large-radius ball around one of the vertices and a small-radius ball around the other vertex contain roughly the same number of elements of $V$.

Consider, however, the following example.

**Example 9 (Tree)** *Consider a complete binary tree, and the shortest-path metric $d(\cdot, \cdot)$ in the tree. Take any two vertices $u$ and $v$ at distance $\frac{1}{2} \log n$. If we look at the ball of radius $d_1$ around $u$ and the ball of radius $d_2 = d_1 + \epsilon \log n$ around $v$, we see that the former has $2^{d_1}$ points in it, and the latter has $2^{d_1} \cdot n^{\epsilon}$ points: it is clearly hopeless to have constant probability of hitting the former and of missing the latter.*

*For every $t < \frac{1}{2} \log n$, however, we have*

$$\mathop{\mathbb{E}}_{A \sim D_t} [|f_A(u) - f_A(v)|] \geq \Omega(1)$$

*because there is a constant probability of hitting one of the $2^{t+1}$ points at distance $\leq t$ from $u$, so that $f_A(u) \leq t$ and also a constant probability of missing the $2^{t+2}$ points*

*at distance $\geq t+1$ from $v$, in which case $f_A(v) \geq t+1$. This is still good, because averaging over all scales we still get*

$$\mathop{\mathbb{E}}_{A\sim D}[|f_A(u) - f_A(v)|] \geq \Omega(1) = \frac{1}{O(\log n)} \cdot d(u,v)$$

*but this example shows that the analysis cannot be restricted to one good scale but, in some cases, we have to take into account the contribution to the expectation coming from all the scales.*

In the above example, the only way to get a ball around $u$ and a ball around $v$ with approximately the same number of points is to get balls of roughly the same radius. No scale could then give a large contribution to the expectation $\mathbb{E}_{A\sim D}[|f_A(u)-f_A(v)|]$; every scale, however, gave a noticeable contribution, and adding them up we had a bounded distortion. The above example will be the template for the full proof, which will do an "amortized analysis" of the contribution to the expectation coming from each scale $t$, by looking at the radii that define a ball around $u$ and a ball around $v$ with approximately $2^t$ elements.

# 2 The Proof of Bourgain's Theorem

Given Fact 2 and Fact 6, proving Bourgain's theorem reduces to proving the following theorem.

**Theorem 10** *For a finite set of points $V$, consider the distribution $D$ over subsets of $V$ sampled by uniformly picking a scale $t \in \{0, \ldots, \log_2 |V|\}$ and then picking independently each $v \in V$ to be in $A$ with probability $2^{-t}$. Let $d : V \times V \to \mathbb{R}$ be a semimetric. Then for every $u, v \in V$,*

$$\mathop{\mathbb{E}}_{A\sim D}[|f_A(u) - f_A(v)|] \geq \frac{1}{c \log_2 |V|} \cdot d(u,v)$$

*where $c$ is an absolute constant.*

Proof: For each $t$, let $ru_t$ be the distance from $u$ to the $2^t$-th closest point to $u$ (counting $u$). That is,

$$|\{w : d(u,w) < ru_t\}| < 2^t$$

and

$$|\{w : d(u,w) \leq ru_t\}| \geq 2^t$$

and define $rv_t$ similarly. Let $t^*$ be the scale such that both $ru_{t^*}$ and $rv_{t^*}$ are smaller than $d(u,v)/3$, but at least one of $ru_{t^*+1}$ or $rv_{t^*+1}$ are $\geq d(u,v)/3$.

Define

$$ru'_t := \min\{ru_t, d(u,v)/3\}$$

and similarly

$$rv'_t := \min\{rv_t, d(u,v)/3\}$$

We claim that there is an absolute constant $c$ such that for every scale $t \in \{0,\ldots,t^*\}$, we have

$$\mathop{\mathbb{E}}_{A\sim D_t} |f_A(u) - f_A(v)| \geq c \cdot (ru'_{t+1} + rv'_{t+1} - ru'_t - rv'_t) \tag{1}$$

We prove the claim by showing that there are two disjoint events, each happening with probability $\geq c$, such that in one event $|f_A(u) - f_A(v)| \geq ru'_{t+1} - rv'_t$, and in the other event $|f_A(u) - f_A(v)| \geq rv'_{t+1} - ru'_t$.

1. The first event is that $A$ avoids the set $\{z : d(u,z) < ru'_{t+1}\}$ and intersects the set $\{z : d(v,z) \leq rv'_t\}$. The former set has size $< 2^{t+1}$, and the latter set has size $\leq 2^t$; the sets are disjoint because we are looking at balls or radius $\leq d(u,v)/3$ around $u$ and $v$; so the event happens with a probability that is at least an absolute constant. When the event happens,

$$|f_A(u) - f_A(v)| \geq f_A(u) - f_A(v) \geq ru'_{t+1} - rv'_t$$

2. The second event is that $A$ avoids the set $\{z : d(v,z) < rv'_{t+1}\}$ and intersects the set $\{z : d(u,z) \leq ru'_t\}$. The former set has size $< 2^{t+1}$, and the latter set has size $\leq 2^t$; the sets are disjoint because we are looking at balls or radius $\leq d(u,v)/3$ around $u$ and $v$; so the event happens with a probability that is at least an absolute constant. When the event happens,

$$|f_A(u) - f_A(v)| \geq f_A(v) - f_A(u) \geq rv'_{t+1} - ru'_t$$

So we have established (1). Averaging over all scales, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{A\sim D} |f_A(u) - f_A(v)| \\
&\geq \frac{c}{1+\log_2 n} \cdot (ru'_{t^*+1} + rv'_{t^*+1} - ru'_0 - rv'_0) \\
&\geq \frac{c}{1+\log_2 n} \cdot \frac{d(u,v)}{3}
\end{aligned}$$

□

There is one remaining point to address. In Fact 2, we proved that a distribution over embeddings on the line can be turned into an L1 embeddings, in which the number of dimensions is equal to the size of the support of the distribution. In our proof, we have used a distribution that ranges over $2^{|V|}$ possible functions, so this would give rise to an embedding that uses a superpolynomial number of dimensions.

To fix this remaining problem, we sample $m = O(\log^3 |V|)$ sets $A_1, \ldots, A_m$ and we define the embedding $f(u) := (m^{-1} \cdot f_{A_1}(u), \ldots, m^{-1} \cdot f_{A_m}(u))$. It remains to prove that this randomized mapping has low distortion with high probability, which is an immediate consequence of the Chernoff bounds. Specifically, we use the following form of the Chernoff bound:

**Lemma 11** *Let $Z_1, \ldots, Z_m$ be independent nonnegative random variables such that, with probability 1, $0 \leq Z_i \leq M$. Let $Z := \frac{1}{m}(Z_1 + \cdots + Z_m)$. Then*

$$\mathbb{P}[\mathbb{E}\, Z - Z \geq t] \leq e^{-2mt^2/M^2}$$

Let us look at any two vertices $u, v$. Clearly, for every choice of $A_1, \ldots, A_m$, we have $||f(u) - f(v)||_1 \leq d(u, v)$ so it remains to prove a lower bound to their L1 distance. Let us call $Z$ the random variable denoting their L1 distance, that is

$$Z := ||f(u) - f(v)|| = \sum_{i=1}^{m} \frac{1}{m} |f_{A_i}(u) - f_{A_i}(v)|$$

We can write $Z = \frac{1}{m} \cdot (Z_1 + \cdots + Z_m)$ where $Z_i := |f_{A_i}(u) - f_{A_i}(v)|$, so that $Z$ is the sum of identically distributed nonnegative random variables, such that

$$Z_i \leq d(u, v)$$
$$\mathbb{E}\, Z_i \geq \frac{c}{\log |V|} d(u, v)$$

Applying the Chernoff bound with $M = d(u, v)$ and $t = \frac{c}{2 \log |V|} d(u, v)$, we have

$$\begin{aligned} &\mathbb{P}\left[Z \leq \frac{c}{2 \log |V|} d(u, v)\right] \\ &\leq \mathbb{P}\left[Z \leq \mathbb{E}\, Z - \frac{c}{2 \log |V|} d(u, v)\right] \\ &\leq 2^{-2mc^2/(2 \log |V|)^2} \end{aligned}$$

which is, say, $\leq 1/|V|^3$ if we choose $m = c' \log^3 |V|$ for an absolute constant $c'$. By taking a union bound over all pairs of vertices,

$$\mathbb{P}\left[\forall u, v.\ ||f(u) - f(v)||_1 \geq \frac{c}{2 \log |V|} \cdot d(u, v)\right] \geq 1 - \frac{1}{|V|}$$