15-859(J): Metric Embeddings and Methods **Lecturer:** Anupam Gupta
Topic: Bourgain's Theorem **Date:** September 11th, 2003
Scribe: Shuchi Chawla and Kedar Dhamdhere

4.1 Introduction

In the last lecture, we studied low-distortion embeddings of general metrics into ℓ_∞ using only a small number of dimensions. In this lecture we will extend that result and give embeddings into ℓ_p for all $1 \le p < \infty$; this fundamental result is essentially due to Bourgain (1985). For completeness, we begin by stating the result from Lecture 3.

Theorem 4.1.1 (Matoušek (1996)) *For any metric space* (X, d)*,*

$$(X, d) \overset{D}{\hookrightarrow} \ell_\infty^{O(Dn^{2/D} \log n)}$$

Theorem 4.1.1 uses the following embedding:

1. Pick random subsets S_{ij}, with $j = 1, 2, \ldots, \frac{D}{2}$ as follows

   At *level* j, for each $i = 1, 2, \ldots, m$ (where $m = O(n^{2/D} \log n)$), form the set S_{ij} by picking each node independently with probability $(n^{-2/D})^j$.

2. Let $f_{ij}(x) = d(x, S_{ij})$ for all $x \in X$. Finally, set

$$f(x) = \oplus_{j=1}^{D/2} \oplus_{i=1}^{m} f_{ij}(x). \tag{4.1.1}$$

An easy application of triangle inequality shows that the above embedding is a contraction. That is, for any two points x and y in X, in every dimension, the distance between x and y is less than $d(x, y)$. This fact will be useful later.

Note that if we use $D = O(\log n)$, we get that $(X, d) \overset{\log n}{\hookrightarrow} \ell_\infty^{O(\log^2 n)}$. Interestingly, we immediately get the fact that the f is a low distortion embedding into ℓ_1 as well, albeit with a worse distortion. Indeed, the contraction, when viewing f as a map into ℓ_1, is at most the contraction with respect to ℓ_∞, which is $O(\log n)$. Moreover, since the embedding is a contraction in every dimension, the *expansion* with respect to ℓ_1 is at most $\log^2 n$, the number of dimensions. Thus we get

$$(X, d) \overset{\log^3 n}{\hookrightarrow} \ell_1^{O(\log^2 n)}$$

Using a more careful analysis, we can show that the above embedding has distortion $O(\log^2 n)$ with a high probability (see Homework 1). As an aside, note that Theorem 4.1.1 is meaningful only for $D = O(\log n)$. For larger values of D, the distortion as well as the number of dimensions is larger that those corresponding to $\log n$.

In the next section we will see the Bourgain's Theorem, which refines the embedding and proof of Theorem 4.1.1, and obtains an embedding into $\ell_1^{O(\log^2 n)}$ with distortion only $O(\log n)$.

4.2 Bourgain's Theorem

Theorem 4.2.1 (Bourgain (1985), Linial, London and Rabinovich (1995)) *For any metric space* (X, d)*, and for any* p*,*

$$(X, d) \xhookrightarrow{O(\log n)} \ell_p^{O(\log^2 n)}$$

Theorem 4.2.1 uses the following embedding, which is very similar to that used for Theorem 4.1.1.

1. Pick random subsets S_{ij}, with $j = 1, 2, \ldots, \log n$ as follows

   At level j, for each $i = 1, 2, \ldots, m$, (with $m = 576 \log n$), form the set S_{ij} by picking each node in X independently with probability 2^{-j}.

   Let $f_{ij}(x) = d(x, S_{ij})$.

2. $f(x) = \oplus_{j=1}^{\log n} \oplus_{i=1}^{m} f_{ij}(x)$.

Let us first prove Theorem 4.2.1 for $p = 1$. Formally, $L = 576 \log n$, and let K denote the number of coordinates, i.e., $K = 576 \, \log^2 n$. We will now prove that

$$(X, d) \xhookrightarrow{64 \log n} \ell_1^K.$$

Lemma 4.2.2 (Expansion) $|f(x) - f(y)|_1 \leq K \, d(x, y)$

Proof: For any set A and any pair of points $x, y \in X$, $|d(x, A) - d(y, A)| \leq d(x, y)$. Since the embedding has K dimensions, the lemma follows. ■

Theorem 4.2.3 (Contraction) *With probability* $\frac{1}{2}$*, for all* $x, y \in X$*,*

$$|f(x) - f(y)| \geq 9 \log n \; d(x, y).$$

Proving this theorem will form a major part of this lecture. We will first describe how it implies Theorem 4.2.1.

Proof of Theorem 4.2.1: By construction, the embedding uses $K \log^2 n$ dimensions. Theorem 4.2.3 shows that f has contraction $\frac{1}{9 \log n}$. On the other hand, the expansion of f is at most $K = 576 \log^2 n$ by Lemma 4.2.2. Multiplying the two, we get that the distortion of the embedding is at most $64 \log n$. ■

Now we begin with a proof of Theorem 4.2.3. Fix a pair of points $x, y \in X$. Let r_j, for $j = 0, 1, \ldots, \log n$, be the smallest radius r such that a ball of radius r around x contains at least 2^j points, that is, $|B(x, r)| \geq 2^j$. Likewise, let r'_j be the smallest radius r for which $|B(y, r)| \geq 2^j$. (Remember that the *ball* $B(x, r)$ is just the set $\{z \in X \,|\, d(x, z) \leq r\}$. Similarly, let the *open ball* $B^o(x, r)$ be defined as $\{z \in X \,|\, d(x, z) < r\}$. Let $\rho_j = \max\{r_j, r'_j\}$. By our construction, we have the following lemma.

Lemma 4.2.4 $|B(x, \rho_j)| \geq 2^j$ *and* $|B(y, \rho_j)| \geq 2^j$*.*

Let t be the smallest value such that $\rho_t + \rho_{t-1} > d(x,y)$; we then redefine $\rho_t = d(x,y) - \rho_{t-1}$. In the following argument, we will consider ρ_j's for values of j at most t.

Remark 4.2.5 The value $\rho_0 = 0$, and the ρ_j's do not decrease as j increases. Furthermore, since $\rho_{t-1} + \rho_t = d(x,y)$ and $\rho_{t-1} \leq \rho_t$, it follows that $\rho_t \geq \frac{1}{2}\, d(x,y)$.

For all j, we define a pair of balls B_j ("bad") and G_j ("good") with the following property. One of these balls is centered around x and the other is centered around y. Moreover, at stage j in the construction of the embedding, there is a reasonable probability that we pick at least one point from G_j and none from B_j. This would give us the necessary separation between x and y in the dimensions corresponding to j. (See Figure 4.2.1). We will formalize this in the rest of the proof.

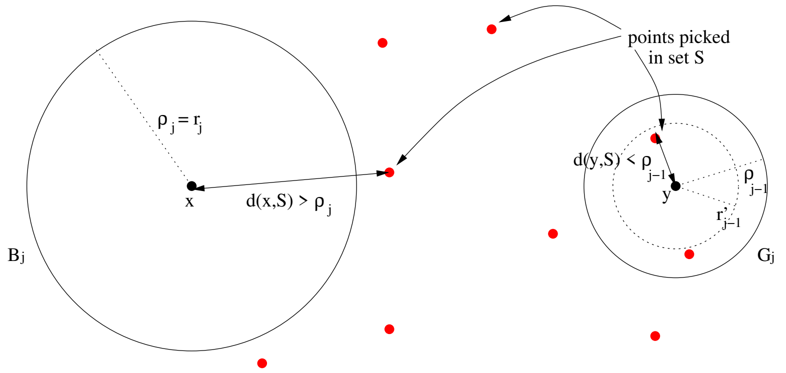

Figure 4.2.1: Balls B_j and G_j for a fixed pair of points x and y and iteration j

If $\rho_j = r_j$, let $B_j = B^o(x, r_j)$ and $G_j = B(y, \rho_{j-1})$. Otherwise, let $B_j = B^o(y, r'_j)$ and $G_j = B(x, \rho_{j-1})$. Note that open ball B_j contains *at most* 2^j points (by definition of r_j and r'_j), and G_j contains *at least* 2^{j-1} points (by Lemma 4.2.4).

Observation 4.2.6 *For all j, $|B_j| \leq 2^j$ and $|G_j| \geq 2^{j-1}$.*

Thus both the balls have size about 2^j. So if we are sampling at rate 2^{-j}, there is a reasonable chance that we will hit the "good" ball and miss the "bad" ball.

Claim 4.2.7 *For any value of j,*

$$\mathbf{Pr}[\textit{ for at least } 18 \log n \textit{ values of } i, |f_{ij}(x) - f_{ij}(y)| \geq (\rho_j - \rho_{j-1})] \geq 1 - \frac{1}{n^3}.$$

Before we prove the claim, let us prove Theorem 4.2.3 using this claim.

Proof of Theorem 4.2.3: Claim 4.2.7 implies that, for a fixed j, we have

$$\mathbf{Pr}\left[\sum_{i=1}^{L} |f_{ij}(x) - f_{ij}(y)| \geq 18 \log n \cdot (\rho_j - \rho_{j-1})\right] \geq 1 - \frac{1}{n^3}$$

In other words, for a fixed j, the probability of not getting "enough" contribution from the L coordinates corresponding to that j is bounded above by $\frac{1}{n^3}$. Since there are at most $\log n$ values of j, the trivial union bound implies that the probability that *any one* of these values of j did not contribute $18 \log n\,(\rho_j - \rho_{j-1})$ to the distance between x and y is $\frac{\log n}{n^3} \leq 1/n^2$. Hence

$$\mathbf{Pr}\left[\sum_{j,i} |f_{ij}(x) - f_{ij}(y)| \geq (18\,\log n)\sum_{j=1}^{t}(\rho_j - \rho_{j-1})\right] \geq 1 - \frac{1}{n^2}.$$

However, note that $\sum_{j,i} |f_{ij}(x) - f_{ij}(y)| = \|f(x) - f(y)\|_1$; on the right hand side, the sum $\sum_{j=1}^{t}(\rho_j - \rho_{j-1})$ telescopes to ρ_t, which is at least $\frac{1}{2}\,d(x,y)$. Hence, for any fixed $x, y \in X$,

$$\mathbf{Pr}\left[\|f(x) - f(y)\|_1 \geq 18\,\log n \frac{d(x,y)}{2}\right] \geq 1 - \frac{1}{n^2}$$

Since there are $\binom{n}{2}$ pairs $x, y \in X$ that we need to argue about, the trivial union bound again implies

$$\mathbf{Pr}\left[\forall x, y \in X\colon \frac{\|f(x) - f(y)\|_1}{d(x,y)} \geq 9\,\log n\right] \geq \frac{1}{2},$$

which proves the theorem. ■

Proof of Claim 4.2.7: Consider the coordinate corresponding to i and j. Recall that we form the set S_{ij} by sampling vertices independently at rate 2^{-j}. Let us consider the case when B_j be centered around x, and G_j around y, as in Figure 4.2.1. (The other case is proved identically.)

We want to compute the probability of the event that we get a good contribution, i.e., of the event

$$\begin{aligned}\mathcal{E} &= \{|f_{ij}(x) - f_{ij}(y)| \geq (\rho_j - \rho_{j-1})\}\\ &= \{|d(S_{ij}, x) - d(S_{ij}, y)| \geq (\rho_j - \rho_{j-1})\}.\end{aligned}$$

It will be tricky to calculate the probability of this event directly, so let us define another event $\mathcal{E}'$ thus

$$\begin{aligned}\mathcal{E}' &= \{d(S_{ij}, x) \geq \rho_j \quad \wedge \quad d(S_{ij}, y) \leq \rho_{j-1}\}\\ &= \{S_{ij} \text{ misses the "bad" ball } B_j \wedge \text{ hits the "good" ball } G_j\ \}.\end{aligned}$$

(Make sure you believe this!) It can be checked that the event $\mathcal{E}'$ implies $\mathcal{E}$, and hence $\mathbf{Pr}[\mathcal{E}] \geq \mathbf{Pr}[\mathcal{E}']$; hence it suffices to lower bound the probability of $\mathcal{E}'$. Finally, let us define

$$\begin{aligned}\mathcal{E}_{\text{hit}} &:= S_{ij} \cap G_j \neq \emptyset\\ \mathcal{E}_{\text{miss}} &:= S_{ij} \cap B_j = \emptyset\end{aligned}$$

and hence $\mathcal{E}' = \mathcal{E}_{\text{hit}} \wedge \mathcal{E}_{\text{miss}}$. Note that $\mathcal{E}_{\text{miss}}$ and $\mathcal{E}_{\text{hit}}$ are independent, since the balls G_j and B_j are disjoint.

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}_{\text{hit}}] &= 1 - \mathbf{Pr}[S_{ij} \cap G_j = \emptyset]\\ &= 1 - (1 - 2^{-j})^{|G_j|}\\ &\geq 1 - (1 - 2^{-j})^{2^{j-1}}\\ &\geq 1 - e^{-1/2} \geq \tfrac{1}{4}\end{aligned}$$

The third step in the above equations follows from the fact that $|G_j| \geq 2^{j-1}$. Likewise, for B_j we have,

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}_{\text{miss}}] &= (1-2^{-j})^{|B_j|} \\ &\geq (1-2^{-j})^{2^j} \geq \tfrac{1}{4}\end{aligned}$$

And finally,

$$\mathbf{Pr}\left[\mathcal{E}'\right] = \mathbf{Pr}[\mathcal{E}_{\text{hit}}] \cdot \mathbf{Pr}[\mathcal{E}_{\text{miss}}] \geq \frac{1}{16}.$$

We now use the following fundamental large-deviations bound, which says that the sum of many independent "well-behaved" random variables is closely concentrated around its expectation. (See, e.g., Alon and Spencer (1992) for a proof.)

Lemma 4.2.8 (Chernoff (1952), Hoeffding (1963)) *Let $X_1, \cdots, X_t$ be binary random variables, with $E[X_i] \geq \mu$ for all $i \leq t$. Let $X = \sum_{i=1}^{t} X_i$. Then, $E[X] \geq \mu t$. The probability that X deviates significantly from its expectation is bounded as follows:*

$$\mathbf{Pr}[X \leq (1-\epsilon)\mu t] \leq e^{-\epsilon^2 \mu t/3}$$

To apply the Chernoff bound, let X_i be the indicator variable that the event $\mathcal{E}'$ happens for S_{ij}. Thus we have $\mathbf{E}[X_i] = \mathbf{Pr}[X_i = 1] \geq \frac{1}{16}$. Hence, by Chernoff bound,

$$\mathbf{Pr}\left[\sum_{i=1}^{L} X_i \geq 18 \log n\right] \geq 1 - \frac{1}{n^3}$$

Thus we have shown that the event $\mathcal{E}'$, and hence $\mathcal{E}$ for at least $18 \log n$ values of i, and hence

$$\mathbf{Pr}[|f_{ij}(x) - f_{ij}(y)| \geq \rho_j - \rho_{j-1} \text{ for at least } 18 \log n \text{ values of } i] \geq 1 - \frac{1}{n^3},$$

proving the claim. ∎

4.3 Embeddings into ℓ_p

In this section, we will show that the same mapping f that we defined for embedding into ℓ_1 works for all $p \geq 1$. It is a fairly surprising result, since we produce a mapping of (X, d) into $\mathbb{R}^{O(\log^2 n)}$ which has a low distortion for all norms ℓ_p *simultaneously.*

Theorem 4.3.1 *There exists an embedding that maps any arbitrary finite metric space (X, d) into $\ell_p^{O(\log^2 n)}$ with $O(\log n)$ distortion.*

$$(X, d) \xhookrightarrow{O(\log n)} \ell_p^{O(\log^2 n)}.$$

In particular, with probability at least $\frac{1}{2}$, the mapping f defined above has this property.

In order to prove this theorem, we will use the Cauchy-Schwarz Inequality.

Fact 4.3.2 (Cauchy-Schwarz) *For two vectors $\vec{a}, \vec{b} \in \mathbb{R}^k$, it holds that $\vec{a}\|_2 \cdot \|\vec{b}\|_2 \geq \langle \vec{a}, \vec{b} \rangle$.*

Proof of Theorem 4.3.1: We will first prove this theorem for $p = 2$. We first give an upper bound on $\|f(x) - f(y)\|_2$. Recall that $K = 576 \log^2 n$ is the number of dimensions of the map f.

$$\begin{aligned}
\|f(x) - f(y)\|_2 &= \Big(\sum_{i,j} |f_{ij}(x) - f_{ij}(y)|^2\Big)^{\frac{1}{2}} \\
&\le \Big(\sum_{i,j} d(x,y)^2\Big)^{\frac{1}{2}} \\
&= \sqrt{K} \cdot d(x,y) \\
&= O(\log n) \cdot d(x,y)
\end{aligned}$$

To give a lower bound on the distance between x and y in the embedding, we use the Cauchy-Schwarz inequality 4.3.2. Set $a_{ij} = |f_{ij}(x) - f_{ij}(y)|$ and $b_{ij} = 1$. Plugging these values in the inequality, we get

$$\begin{aligned}
\Big(\textstyle\sum_{i,j} |f_{i,j}(x) - f_{i,j}(y)|^2\Big)^{\frac{1}{2}} \cdot \sqrt{K} &\ge \sum_{i,j} |f_{ij}(x) - f_{ij}(y)| \\
\sqrt{K} \cdot \|f(x) - f(y)\|_2 &\ge \|f(x) - f(y)\|_1 \\
&\ge O(\log n) \cdot d(x,y),
\end{aligned}$$

the last inequality following from Theorem 4.2.1. Now using that $\sqrt{K} = \Theta(\log n)$, we get that

$$\|f(x) - f(y)\|_2 \ge \frac{d(x,y)}{O(1)},$$

which proves the theorem. ■

This result can be easily generalized to ℓ_p using the Hölder's inequality (see Homework 2).

Fact 4.3.3 (Hölder's Inequality) *For two vectors $\vec{a}, \vec{b} \in \mathbb{R}^k$, and $p, q \in \mathbb{R}^+$ such that $\frac{1}{p} + \frac{1}{q} = 1$, it holds that $\|\vec{a}\|_p \cdot \|\vec{b}\|_q \ge \langle \vec{a}, \vec{b} \rangle$.*

Remark 4.3.4 As a sanity check, notice that if $p = q = 2$, then this gives us the Cauchy-Schwarz inequality from Fact 4.3.2. Furthermore, if $p = 1$ and $q = \infty$, then the inequality implies that $\|\vec{a}\|_1 \cdot b_{\max} \ge \langle \vec{a}, \vec{b} \rangle$, which can also be easily checked.

A stronger version of the theorem 4.3.1 proved by Matoušek (1997) gives the following result:

$$(X, d) \xhookrightarrow{O(\log n/p)} \ell_p$$

4.4 Optimal embedding into ℓ_2 via SDP

In the previous section, we saw that any metric space embeds into the Euclidean space with $O(\log n)$ distortion. One can ask the following questions.

Question 1: Given a metric (X, d), can we check in polynomial time if $(X, d) \xhookrightarrow{1} \ell_2$?

Furthermore, Theorem 4.2.1 implies that $D \le O(\log n)$, but it could be much smaller for a specific metric space. This motivates the question:

Question 2: If a metric space (X, d) embeds into ℓ_2 with distortion D, i.e. $(X, d) \stackrel{D}{\hookrightarrow} \ell_2$, can we compute D in polynomial time? Failing that, can we approximate D?

Answer 1: Let $X = \{x_1, x_2, \ldots, x_n\}$. Let f denote the mapping from X to ℓ_2. Let $f(x_i) = v_i$. By suitable change of coordinates, we can assume that $f(x_1) = v_1 = \vec{0}$. Now answering Question 1 in the affirmative is equivalent to checking if

$$\begin{aligned} & & \|v_i - v_j\|^2 &= d_{ij}^2 & \forall i, j \\ \Leftrightarrow & & \|v_i\|^2 + \|v_j\|^2 - 2\langle v_i, v_j\rangle &= d_{ij}^2 & \\ \Leftrightarrow & & \langle v_i, v_j\rangle &= \tfrac{1}{2}(d_{1i}^2 + d_{1j}^2 - d_{ij}^2), & \end{aligned}$$

where the last line follows from the fact that, for all i, $\|v_i\|^2 = \|v_i - v_1\|^2 = d_{1i}^2$.

Imagine a matrix A, with $A_{ij} = \frac{1}{2}(d_{1i}^2 + d_{1j}^2 - d_{ij}^2)$; the above argument implies that we want $A_{ij} = \langle v_i, v_j\rangle$. However, this is equivalent to asking whether A can be written as $B^T B$ for the

$$B = \begin{pmatrix} \cdots & v_1 & \cdots \\ & \vdots & \\ \cdots & v_n & \cdots \end{pmatrix}.$$

It is known that such a decomposition exists if and only if A is a *positive semi-definite matrix*, and that it can be obtained in polynomial time; see, e.g., page 295 of Grötschel, Lovász, and Schrijver (1988). ■

Answer 2: The basic ideas in the previous answer can be used to answer Question 2 as well. Formally, we will write a *Semi-Definite Program* (SDP) to obtain the value of the best distortion. One way of writing an SDP is as a linear program, where the "variables" are inner products of vectors.

Example: Here is an example of a semi-definite program for the Max-Cut problem:

$$\begin{aligned} \min \quad & \sum_{i,j} w_{ij}\left(\frac{1 - \langle v_i, v_j\rangle}{2}\right) & \\ \text{s.t.} \quad & \langle v_i, v_i\rangle = 1 & \forall\, i \\ & \langle v_i, v_j\rangle + \langle v_j, v_k\rangle + \langle v_k, v_i\rangle \geq -1 & \forall\, i, j, k \end{aligned}$$

■

In our problem, we want to minimize D, subject to the constraints that all the pairwise distances are distorted by atmost D. To formulate it as an SDP, we change D to a vector whose norm will denote the distortion.

$$\begin{aligned} \min \quad & \langle \vec{D}, \vec{D}\rangle & \\ \text{s.t.} \quad & d_{ij}^2 \leq \langle v_i, v_i\rangle + \langle v_j, v_j\rangle - 2\langle v_i, v_j\rangle \leq \langle \vec{D}, \vec{D}\rangle d_{ij}^2 & \forall\, i, j \\ & d_{1i}^2 \leq \langle v_i, v_i\rangle \leq \langle \vec{D}, \vec{D}\rangle d_{1i}^2 & \forall\, i \end{aligned}$$

It is a well known fact that, there is a polynomial time algorithm to solve an SDP to any desired precision. Thus we have an algorithm which finds the minimum possible distortion for embedding an arbitrary metric space (X, d) into the Euclidean space. ■

4.4.1 Problems in ℓ_1-embeddability

The analogous questions can also be asked about embeddability into ℓ_1:

Question 1': Given a metric (X, d), can we check in polynomial time if $(X, d) \overset{1}{\hookrightarrow} \ell_1$, i.e., if it embeds isometrically into ℓ_1?

However, it is known that this problem is NP-hard, as was shown by Karzanov (1985) via a reduction from MAX-CUT. On the other hand, little is known about the following problem:

Question 2': Given a metric (X, d), approximate the value of the best distortion D such that $(X, d) \overset{D}{\hookrightarrow} \ell_1$.

Since $D \geq 1$ for all metrics, Bourgain's theorem implies a $O(\log n)$ approximation algorithm, and this is the best known. In fact, even if we are given a metric that is *isometrically* ℓ_1 embeddable, we have no algorithms that can guarantee a distortion of $o(\log n)$.

References

Alon, N. and Spencer, J. (1992). *The Probabilistic Method.* Wiley Interscience, New York.

Bourgain, J. (1985). On Lipschitz embeddings of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1-2):46–52.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507.

Grötschel, M., Lovász, L., and Schrijver, A. (1988). *Geometric algorithms and combinatorial optimization.* Springer-Verlag, Berlin.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30.

Karzanov, A. V. (1985). Metrics and undirected cuts. *Mathematical Programming*, 32(2):183–198.

Linial, N., London, E., and Rabinovich, Y. (1995). The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245. (Preliminary version in *35th FOCS*, 1994).

Matoušek, J. (1996). On the distortion required for embedding finite metric spaces into normed spaces. *Israel J. Math.*, 93:333–344.

Matoušek, J. (1997). On embedding expanders into l_p spaces. *Israel J. Math.*, 102:189–197.