Seminar on Algorithms and Geometry

Lecture 6

June 4, 2009

Lecturer: Robert Krauthgamer Scribe by: Inbal Talgam Updated: June 14, 2009

A Communication Complexity Perspective on Metric Spaces

Today we study the distance estimation problem in ℓ_1 from a communication complexity perspective. Along the way we shall find a (weak) analogue in ℓ_1 for the Johnson-Lindenstrauss dimension reduction lemma.

1 Setting

Assume Alice and Bob each have a private input of size n. They need to exchange enough information to be able to calculate a function of both values. We seek protocols that use the least possible communication in terms of:

1. Number of bits exchanged.
2. Number of communication rounds.

We focus on randomized protocols which are required to succeed with probability (say) $\geq \frac{2}{3}$ over public random coins. Another requirement is *simultaneousness* - Alice and Bob send only one message each to a referee, who calculates the output. Note that a simultaneous protocol is a particular case of a 1-round protocol (the referee can be simulated by one of the players, who is thus able to calculate the output after a single round).

2 Randomized Simultaneous Protocol for Equality Testing

We illustrate the concept of protocols with the problem of equality testing. We will use this protocol later.

Private inputs: $x, y \in \{0,1\}^n$.

Output: Accept iff $x = y$.

Protocol: Alice and Bob choose $r \in \{0,1\}^n$ at random and send the referee $\langle x,r\rangle, \langle y,r\rangle$, where $\langle x,r\rangle = \sum_{i=1}^{n} x_i r_i$ (all calculations are modulo 2). The referee accepts iff $\langle x,r\rangle = \langle y,r\rangle$.

Analysis: If $x = y$ then $\Pr[accept] = 1$. If $x \neq y$ then there exists an index j such that $x_j \neq y_j$. The protocol accepts iff:

$$\begin{aligned} 0 = \langle x,r\rangle + \langle y,r\rangle &= \\ \sum_{i=1}^{n} (x_i + y_i) r_i &= \\ \sum_{i \neq j} (x_i + y_i) r_i + (x_j + y_j) r_j &= \\ \sum_{i \neq j} (x_i + y_i) r_i + 1 \cdot r_j & \end{aligned}$$

By independence of the r_i's this happens exactly with probability $\frac{1}{2}$. This probability can be lowered by repeating the protocol.

3 The Distance Estimation Problem in ℓ_1

Private inputs: $x, y \in \ell_1$. Wlog we assume $x, y \in \{0,1\}^m$.

Output: For an approximation parameter $\alpha \geq 1$ and a threshold parameter $R > 0$, decide whether $\|x-y\|_1 \leq R$ or $\|x-y\|_1 > \alpha R$ (this is the decision version of the distance estimation within factor α).

Theorem 1 *[Kushilevitz, Ostrovsky & Rabani, 2000] For every $0 < \epsilon < 1$ and $R > 0$, there is a randomized simultaneous protocol for estimating the ℓ_1-distance within factor $\alpha = 1+\epsilon$ using $O\left(\frac{1}{\epsilon^2}\right)$ bits of communication.*

Proof

Protocol: Alice and Bob choose $r \in \{0,1\}^n$ such that $r_i = 1$ with probability $\frac{1}{2R}$ and otherwise $r_i = 0$. As before they send the referee $\langle x,r\rangle, \langle y,r\rangle$. This is repeated $T = O\left(\frac{1}{\epsilon^2}\right)$ times. The referee accepts iff $\langle x,r\rangle = \langle y,r\rangle$ in at least β-fraction of the T repetitions.

Analysis: We can think of r_i as if it were chosen in 2 independent steps: First, a random subset $S \subseteq \{1,\ldots,n\}$ is selected by including each i independently with probability $\frac{1}{R}$; Then, if $i \notin S$ then r_i is set to 0, and if $i \in S$ then r_i is set to be a fair coin, i.e. 1 with probability $\frac{1}{2}$ and 0 otherwise. Denote by x_S (similarly y_S) the restriction of x to the positions in S. Once S is selected, the probability to accept is exactly as in example 1:

$$\begin{aligned} \Pr[\langle x,r\rangle = \langle y,r\rangle \,|S] &= \\ \Pr[\langle x_S,r_S\rangle = \langle y_S,r_S\rangle \,|S] &= \begin{cases} 1 & x_S = y_S \\ \frac{1}{2} & o.w. \end{cases} \end{aligned}$$

By the law of total probability:

$$\Pr\left[\langle x,r\rangle = \langle y,r\rangle\right] = \frac{1}{2}\Pr\left[x_S = y_S\right] + \frac{1}{2}$$

Notice that $\Pr\left[x_S = y_S\right] = \left(1-\frac{1}{R}\right)^{\|x-y\|_1}$, and so:

$$\|x-y\|_1 \le R \implies P_{YES} := \Pr\left[x_S = y_S\right] \ge \left(1-\frac{1}{R}\right)^R$$

$$\|x-y\|_1 > (1+\epsilon)R \implies P_{NO} := \Pr\left[x_S = y_S\right] < \left(1-\frac{1}{R}\right)^{(1+\epsilon)R}$$

An easy calculation shows that $P_{YES} - P_{NO} = \Omega(\epsilon)$ (using the fact that $P_{YES} \le e^{-1}$ and the bound $e^{-\epsilon} \le 1-\epsilon+\frac{\epsilon^2}{2}$, and assuming wlog $R \ge 2$). Since we repeat everything $T = O\left(\frac{1}{\epsilon^2}\right)$ times, with high probability the number of *accepts* will be concentrated around its expectation, which is $T\left(\frac{1}{2}P_{YES}+\frac{1}{2}\right)$ if $\|x-y\|_1 \le R$ and $T\left(\frac{1}{2}P_{NO}+\frac{1}{2}\right)$ if $\|x-y\|_1 > (1+\epsilon)R$. Therefore we choose β to be "in the middle" between the 2 expectations, i.e. $\beta = \frac{1}{2} + \frac{P_{YES}+P_{NO}}{4}$. A success probability of $\ge \frac{2}{3}$ can now be proved using Chernoff's bound. ■

Corollary 2 *Given n points $x_1,\ldots,x_n \in \ell_1$ and parameters $0<\epsilon<1$, $R>0$, there is a map $f:\ell_1 \to \{0,1\}^T$ for $T = O\left(\frac{1}{\epsilon^2}\log n\right)$ such that for all i,j:*

$$\|x_i - x_j\|_1 \le R \implies \|f(x_i) - f(x_j)\|_1 \le \beta T$$

$$\|x_i - x_j\|_1 > (1+\epsilon)R \implies \|f(x_i) - f(x_j)\|_1 > \beta T$$

In fact, the map f is constructed at random independently of the points, and thus we can derive a Near Neighbor Search algorithm for ℓ_1^d with approximation $1+\epsilon$, query time $O\left(\frac{1}{\epsilon^2}\log n + d\right)$ and preprocessing $dn^{O\left(\frac{1}{\epsilon^2}\right)}$, just by preparing in advance answers for all queries.

See Handout 7 for a lower bound on communication complexity and for research directions.