22. The Gap-Hamming Problem

Lecturer: Meena Mahajan *Scribe: Abhishek Dang*

In this and the next class, we examine the Gap Hamming problem and show that it requires $\Omega(n)$ communication.

22.1 Problem Statement

Given $x, y \in \{-1, 1\}^n$ with the promise that $|\langle x, y\rangle| \geq \sqrt{n}$, decide whether $\langle x, y\rangle \leq -\sqrt{n}$ or $\geq \sqrt{n}$. (If the input does not satisfy the promise, any outcome is acceptable.) That is

$$\mathsf{GHD}_n(x, y) = \begin{cases} -1 & \text{if } \langle x, y\rangle \leq -\sqrt{n} \\ 1 & \text{if } \langle x, y\rangle \geq \sqrt{n} \\ \text{anything} & \text{otherwise} \end{cases}$$

The following reformulation of the problem justifies the nomenclature. Denoting Hamming distance (the number of coordinates where x, y disagree) by $\Delta(x, y)$, we can see that

$$\langle x, y\rangle = n - 2\Delta\langle x, y\rangle$$

Thus,

$$|\langle x, y\rangle| \geq \sqrt{n} \iff |\Delta(x, y) - \frac{n}{2}| \geq \frac{\sqrt{n}}{2}$$

So the problem can be reformulated as follows: Given strings x, y satisfying the promise that the Hamming distance between them is not very close to (not within $\sqrt{n}/2$ of) $n/2$, decide whether it exceeds or is less than $n/2$.

$$\mathsf{GHD}_n(x, y) = \begin{cases} -1 & \text{if } \Delta(x, y) \geq \frac{n}{2} + \frac{\sqrt{n}}{2} \\ 1 & \text{if } \Delta(x, y) \leq \frac{n}{2} - \frac{\sqrt{n}}{2} \end{cases}$$

with the indicated promise.

Generalized problem statement

More generally, we may consider any threshold t and any gap g. Under the promise that $\langle x, y\rangle$ is not within $\pm g$ of t, decide whether or not it exceeds t. Denote this generalised problem on length n strings by $\mathsf{GHD}_{n,t,g}$.

$$\mathsf{GHD}_{n,t,g}(x, y) = \begin{cases} -1 & \text{if } \langle x, y\rangle \leq t - g \\ +1 & \text{if } \langle x, y\rangle \geq t + g \end{cases}$$

With this notation, the problem we introduced above is $\mathsf{GHD}_{n,0,\sqrt{n}}$.

While the general problem is of interest, we show that it reduces to GHD with only a polynomial blowup in n.

Claim 22.1. *$GHD_{n,t,g}$ reduces to $GHD_{n+|t|,0,g}$.*

Proof. We simply pad the given strings with $k = |t|$ additional bits, as follows.

$$x \longrightarrow X = \boxed{x \mid -1 \ldots -1}$$
$$y \longrightarrow Y = \boxed{y \mid \pm 1 \ldots \pm 1}$$

where we choose to use one of the signs uniformly in the latter padding. Hence $\langle X, Y\rangle = \langle x, y\rangle - k$ or $\langle x, y\rangle + k$ depending on this choice. We pad y with -1s when $t < 0$ and with $+1$s when $t > 0$. In either case we have $\langle X, Y\rangle = \langle x, y\rangle - t$, giving a reduction to $\mathrm{GHD}_{n+|t|,0,g}$. □

Claim 22.2. *$\mathsf{GHD}_{n,0,g}$ reduces to $\mathsf{GHD}_{N,0,\sqrt{N}}$ for a suitable $N \le n^2/g^2$.*

Proof. This reduction is also achieved by padding, but this time by a number of copies of the input.

If $g \ge \sqrt{n}$, we simply use the same inputs for the $\mathsf{GHD}_{n,0,\sqrt{n}}$ problem. (So $N = n$.)

Otherwise consider the map

$$x \longrightarrow X = \overbrace{x \ldots x}^{k}; \qquad y \longrightarrow Y = \overbrace{y \ldots y}^{k}.$$

This amplifies the gap, because $\langle X, Y\rangle = k\langle x, y\rangle$. We would be done if $kg \ge \sqrt{N} = \sqrt{kn}$, so we choose $k = n/g^2$ (and $N = n^2/g^2$). □

Henceforth, we restrict attention to the special case $\mathsf{GHD}_{n,0,\sqrt{n}}$ and denote it by GHD_n or just GHD.

22.2 Motivation

The gap-Hamming distance problem (GHD) was proposed by Indyk and Woodruff [IW05] as a means to understand the complexity of several streaming tasks. We would be interested in the applications of GHD to computing frequency moments of data streams. More specifically, we briefly study how lower bounds on the communication complexity of gap-Hamming distance imply lower bounds on the memory requirements of estimating the number of distinct elements in a data stream (denoted F_0). (Recall the discussion in Lecture 5.)

For any stream of n numbers from $[m]$, we know that F_0 can be found by a 1-pass algorithm using $O(m)$ space, and can be aproximated by a 1-pass algorithm using $O(\log m, \log n)$ space ([AMS99]). Further, ignoring polylog m factors, it was shown in [BJKS04] that a 1-pass algorithm can ε-approximate F_0 in space $O(1/\varepsilon^2)$. A natural question to ask is whether the dependence on ε can be improved. Indyk and Woodruff showed in [IW05] that it cannot; a quadratic dependence on $1/\varepsilon$ is necessary. We will see the proof of this result.

22.3 Streaming lower bounds from GHD lower bounds

Theorem 22.3 (Indyk,Woodruff [IW05]). *Suppose there exists a randomized 1-pass algorithm $\mathcal{A}$ that uses space S, and with probability at least 2/3, estimates F_0 within error at most ε. Then there exists a 1-way communication protocol for GHD_M with $M = \lfloor 1/(16\varepsilon^2) \rfloor$, with error less than or equal to $1/3$, and complexity $S + O(\log 1/\varepsilon)$.*

Proof. For the purpose of this proof, we use the Hamming distance formulation of the GHD_M problem.

The algorithm $\mathcal{A}$, on an input stream, uses some randomness r and outputs a number $\widetilde{F_0}$. The algorithm satisfies the following:

$$\Pr_r[|\widetilde{F_0} - F_0| > \varepsilon F_0] \leq \frac{1}{3}$$

Consider an instance of GHD_M where Alice and Bob are given strings $x, y \in \{0,1\}^M$ respectively. Alice constructs from x a stream (or set) a_x of numbers in $[M]$, where $i \in a_x$ iff $x_i = 1$. Similarly Bob constructs from y a stream a_y. Alice runs $\mathcal{A}$ on a_x and sends the resulting state of $\mathcal{A}$ to Bob. Furthermore, she sends across the number of 1's in x. Bob resumes $\mathcal{A}$ on a_y, thus computing an estimate $\widetilde{F_0}$ for F_0. Bob accepts if

$$2\widetilde{F_0} - |x|_1 - |y|_1 > \frac{M}{2}.$$

Cost: Alice needs to send S bits for the state of $\mathcal{A}$, and $\lceil \log M \rceil$ bits for the number of 1's in x. So this protocol has cost $O(S + \log(1/\varepsilon))$.

Analysis: Denote the cardinalities of the sets $a_x \setminus a_y$, $a_x \cap a_y$, and $a_y \setminus a_x$ by a, b, c respectively. We write,

$$\begin{aligned}
&F_0 = a + b + c \\
&H = \Delta(x,y) = a + c \\
&H \leq F_0 \leq M \\
&a_x = |x|_1 = a + b, \quad a_y = |y|_1 = b + c \\
&H = 2F_0 - |x|_1 - |y|_1 = 2F_0 - W && \text{(say)} \\
&E := 2\widetilde{F_0} - |x|_1 - |y|_1 = 2\widetilde{F_0} - W \\
&(1-\varepsilon)F_0 \leq \widetilde{F_0} \leq (1+\varepsilon)F_0 && \text{(by assumption about } \mathcal{A}) \\
&H - 2\varepsilon F_0 \leq E \leq H + 2\varepsilon F_0
\end{aligned}$$

Case 1: $H > \frac{M}{2} + \frac{\sqrt{M}}{2}$. Then, since $F_0 \leq M$,

$$E \geq H - 2\varepsilon F_0 > \left(\frac{M}{2} + \frac{\sqrt{M}}{2}\right) - 2\varepsilon M$$

We want $E > M/2$ in this case. So we need $\frac{\sqrt{M}}{2} \geq 2\varepsilon M$.

Case 2: $H < \frac{M}{2} - \frac{\sqrt{M}}{2}$. Then

$$E \leq H + 2\varepsilon F_0 < \left(\frac{M}{2} - \frac{\sqrt{M}}{2}\right) + 2\varepsilon M$$

We want $E < M/2$ in this case. So again we need $\frac{\sqrt{M}}{2} \geq 2\varepsilon M$.

So, the theorem holds for $\frac{\sqrt{M}}{2} \geq 2\varepsilon M$, that is, $M \leq \frac{1}{16\varepsilon^2}$. □

Corollary 22.4. *If the randomized one-way communication complexity of* $\mathrm{GHD_M}$ *satisfies*

$$\mathrm{R}^{1-way}(\mathrm{GHD_M}) = \Omega(M)$$

then any ε-approximate 1-pass streaming algorithm for F_0 needs $\Omega(1/\varepsilon^2)$ space.

22.4 1-way lower bound for GHD

We now show the 1-way lower bound for GHD. This, along with Corollary 22.4, tells us that 1-pass streaming algorithms need $\Omega(1/\varepsilon^2)$ space to approximate F_0. We follow the proof from [JKS08].

Theorem 22.5 ([Woo04, JKS08])**.**

$$\mathrm{R}^{1\text{-}way}(\mathsf{GHD}_N) = \Omega(N)$$

Proof. We give a reduction from the INDEX problem. As has been seen earlier in the course (Lectures 9,10),

$$\mathrm{R}^{\text{1-way}}_{1/2-\delta}(\text{INDEX}) \geq 2(\log e)\delta^2 n = \Omega(n)$$

Say we have a protocol Π for GHD; we show how this can also be used to solve INDEX.

Consider an instance of INDEX, with Alice having as input $x \in \{0,1\}^n$ and Bob an index $i \in [n]$. We consider only the case when n is odd; the other case naturally reduces to this one, and we invoke this assumption in the analysis.

Alice and Bob use public coins to choose an $N \times n$ matrix R with ± 1 entries (we will specify N a bit later). Let $(R_j)_{j=1}^N$ denote the rows of R. Alice (naturally) changes her $x \in \{0,1\}^n$ to $X \in \{1,-1\}^n$ and then further computes $Y = (Y_1, \ldots, Y_N)$ where $Y_j = \mathsf{sgn}(\langle X, R_j \rangle)$. (Recall that n is odd, so $\langle X, R_j \rangle \neq 0$. We use the convention that $\mathsf{sgn}(a) = 1$ if $a > 0$, -1 otherwise.) Bob picks out the i^{th} column of R; denote this column by Z. Alice and Bob then run the GHD protocol Π on Y, Z, and report the same outcome (mapping back ± 1 to $\{0,1\}$) for the given instance of INDEX.

Let us analyze the probability that this protocol answers the INDEX instance correctly.

Let r denote a row of R. Then r is a uniformly random element of $\{\pm 1\}^n$. Notice that r contributes to $\Delta(Y,Z)$ exactly when $\mathsf{sgn}\langle X, r\rangle \neq r_i$. That is,

$$\Delta(Y,Z) = |\{j \mid \mathsf{sgn}(\langle X, R_j \rangle \neq R_{ji})\}|$$

We will show below the following two claims.

Claim 22.6. *There exists a constant $c > 0$ such that for every $X \in \{\pm 1\}^n$ and every $i \in [n]$,*

$$\Pr_{r \in \{\pm 1\}^n}[\mathsf{sgn}(\langle X, r\rangle) \neq r_i] \quad \begin{cases} \geq & \frac{1}{2} + \frac{c}{\sqrt{n}} \quad \text{if } X_i = -1 \\ \leq & \frac{1}{2} - \frac{c}{\sqrt{n}} \quad \text{if } X_i = +1 \end{cases}$$

Claim 22.7. *For a suitable chosen $N \in \Theta(n)$, for every $X \in \{\pm 1\}^n$ and every $i \in [n]$, for Y, Z set as above,*

$$\Pr_{R \in \{\pm 1\}^{N \times n}}\left[\Delta(Y,Z) \geq \frac{N}{2} + \sqrt{N}\right] \quad \geq \frac{2}{3} \quad \text{if } X_i = -1$$

$$\Pr_{R \in \{\pm 1\}^{N \times n}}\left[\Delta(Y,Z) \leq \frac{N}{2} - \sqrt{N}\right] \quad \geq \frac{2}{3} \quad \text{if } X_i = +1$$

Assuming these claims, now we can conclude that for every input x, i to INDEX, with probability at least $2/3$ over the choice of R, $\mathsf{GHD}_M(Y,Z)$ correctly answers $\mathrm{INDEX}(x,i)$.

The protocol we have designed for INDEX is also one-way. Its cost is the same as the cost of Π. Hence

$$\Omega(n) \leq \mathrm{R}^{\text{1-way}}(\mathrm{INDEX}_n) \leq \mathrm{R}^{\text{1-way}}(\mathsf{GHD}_N)$$

□

We now prove the two claims.

Proof. (Of Claim 22.6) Let us write

$$\langle X, r\rangle = X_i r_i + w \quad \text{where } w = \sum_{j \neq i} X_j r_j$$

Remember that we had assumed n to be odd, so $w \neq 0 \implies |w| \geq 2$. Thus if $w \neq 0$, then the first term $X_i r_i$ is irrelevant in deciding $\mathsf{sgn}\langle X, r\rangle$. We use this fact. Denoting $\Pr_r[w = 0]$ by α we have

$$\begin{aligned}
& \Pr_r[\mathsf{sgn}(\langle X, r\rangle) \neq r_i] \\
= \; & \Pr_r[\mathsf{sgn}(X_i r_i + w) \neq r_i] \\
= \; & \Pr_r[\mathsf{sgn}(X_i r_i + w) \neq r_i | w \neq 0] \cdot Pr_r[w \neq 0] + \Pr_r[\mathsf{sgn}(X_i r_i + w) \neq r_i | w = 0] \cdot Pr_r[w = 0] \\
= \; & \Pr_r[\mathsf{sgn}(w) \neq r_i] \cdot (1 - \alpha) + \Pr_r[X_i = -1] \cdot \alpha \\
= \; & \frac{1}{2} \cdot (1 - \alpha) + \Pr_r[X_i = -1] \cdot \alpha \quad \text{because } w \text{ and } r_i \text{ are independent}
\end{aligned}$$

Since X_i is a fixed bit independent of r, $\Pr_r[X_i = -1]$ is either 0 or 1. Thus we see that

$$\Pr_r[\mathsf{sgn}\langle X, r\rangle \neq r_i] \quad = \quad \begin{cases} \frac{1}{2} \cdot (1 - \alpha) & \text{if } X_i = +1 \\ \frac{1}{2} \cdot (1 - \alpha) + \alpha = \frac{1}{2} \cdot (1 + \alpha) & \text{if } X_i = -1 \end{cases}$$

Now using the definition of α and Stirling's approximation, we see that for some absolute constant c',

$$\alpha = \Pr_r[w = 0] = \Pr_r[\sum_{j \neq i} X_j r_j = 0] = \frac{\binom{n-1}{(n-1)/2}}{2^{n-1}} \sim \frac{c'}{\sqrt{n}}$$

Choosing $c = c'/2$, we obtain the claimed statement. □

Proof. (Of Claim 22.7) By Claim 22.6, we know bounds on the probability with which row R_j contributes to $\Delta(Y, Z)$. Since there are N rows, using linearity of expectation, we have

$$\mathbb{E}[\Delta(Y,Z)] \begin{cases} \leq \frac{N}{2} - \frac{cN}{\sqrt{n}} & \text{if } X_i = 1 \\ \geq \frac{N}{2} + \frac{cN}{\sqrt{n}} & \text{if } X_i = -1 \end{cases}$$

Also, the rows are all independent, So using Chernoff's bounds with $N = \left(\frac{2\sqrt{n}}{c}\right)^2 = \frac{4n}{c^2}$ gives the claimed statement.

Applying Chernoff Bound: Details. Let W_j be a 0-1 indicator variable for whether the jth row contributes to $\Delta(Y, Z)$, and let $W = \sum_j W_j = \Delta(Y, Z)$. Claim 22.6 gives bounds on the probability that each $W_j = 1$, depending on whether or not $X_i = 1$. Now from Chernoff bound,

$$\Pr[|W - \mathbb{E}[W]| > \varepsilon] \leq e^{-2\varepsilon^2/N}$$

We choose $N = \left(\frac{2\sqrt{n}}{c}\right)^2 = \frac{4n}{c^2}$.

If $X_i = -1$, then $\mathbb{E}[W] \geq \frac{N}{2} + \frac{cN}{\sqrt{n}} = \frac{N}{2} + 2\sqrt{N}$. So

$$\Pr[W < \frac{N}{2} + \sqrt{N}] \leq \Pr[|W - \mathbb{E}[W]| > \sqrt{N}] \leq e^{-2} < 1/3.$$

Similarly, if $X_i = 1$, then $\mathbb{E}[W] \leq \frac{N}{2} - \frac{cN}{\sqrt{n}} = \frac{N}{2} - 2\sqrt{N}$. So

$$\Pr[W > \frac{N}{2} - \sqrt{N}] \leq \Pr[|W - \mathbb{E}[W]| > \sqrt{N}] \leq e^{-2} < 1/3.$$

□

References

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. *The space complexity of approximating the frequency moments.* J. Computer and System Sciences, 58(1):137–147, 1999. (Preliminary Version in *28th STOC*, 1996). doi:10.1006/jcss.1997.1545.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. *An information statistics approach to data stream and communication complexity.* J. Computer and System Sciences, 68(4):702–732, June 2004. (Preliminary Version in *43rd FOCS*, 2002). doi:10.1016/j.jcss.2003.11.006.

[IW05] Piotr Indyk and David P. Woodruff. *Optimal approximations of the frequency moments of data streams.* In *Proc. 37th ACM Symp. on Theory of Computing (STOC)*, pages 202–208. 2005. doi:10.1145/1060590.1060621.

[JKS08] T. S. Jayram, Ravi Kumar, and D. Sivakumar. *The one-way communication complexity of Hamming distance.* Theory of Computing, 4(1):129–135, 2008. doi:10.4086/toc.2008.v004a006.

[Woo04] David P. Woodruff. *Optimal space lower bounds for all frequency moments.* In *Proc. 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175. 2004. doi:10.1145/982792.982817.