Tight Lower Bounds for Selection in Randomly Ordered Streams

(Extended Abstract)

Amit Chakrabarti*
ac@cs.dartmouth.edu

T. S. Jayram
jayram@almaden.ibm.com

Mihai Pătraşcu*
mip@mit.edu

Abstract

We show that any algorithm computing the median of a stream presented in random order, using $\text{polylog}(n)$ space, requires an optimal $\Omega(\log \log n)$ passes, resolving an open question from the seminal paper on streaming by Munro and Paterson, from FOCS 1978.

1 Introduction

Finding order statistics in a data stream has been studied since the classic work of Munro and Paterson from FOCS 1978 [MP80], which is often cited as a paper introducing the streaming model. Since then, much ink has been spilled over this problem [MRL98, MRL99, GK01, GKMS02, GZ03, CM05, SBAS04, CKMS06, GM06, GM07]. Indeed, this is likely one of the most studied problems in the streaming model, comparable only to the question of estimating frequency moments. Order statistics (a.k.a. quantiles, percentiles, median) are one of the most natural and frequently used summaries of a data set, making the streaming problem arise constantly in practical settings, such as large databases. (Indeed, most of the works referenced have appeared in "practical" conferences.)

Munro and Paterson [MP80] study the problem under two assumptions about the stream order: adversarial and random order. In the case of random order, any permutation of the input values is equally likely to appear in the stream. However, multiple passes through the stream read items in the same order. They give algorithms that make p passes, and require $\widetilde{O}(n^{1/p})$ memory[1] for adversarial order, and $\widetilde{O}(n^{1/2p})$ memory for random order. They also prove some lower bounds under assumptions about what the algorithm can store in its internal memory. In the adversarial model, an unrestricted tight lower bound of $\widetilde{\Omega}(n^{1/p})$ was shown by Guha and McGregor [GM07]. Their bounds are also tight for the well-studied problem of approximate selection, where given a rank r one must return a value of rank $r \pm \Delta$.

The random order model is highlighted in the conclusion of [MP80] as an important challenge for future research. Munro and Paterson conjecture that selection with $\text{polylog}(n)$ space requires $\Theta(\log \log n)$ passes. In PODS 2006, Guha and McGregor [GM06] show the upper-bound side of this conjecture. In general, they show that an algorithm using p passes can return an element with rank $r \pm \widetilde{O}(n^{2^{-p}})$. It follows immediately that with $O(p)$ passes one can perform *exact* selection using space $\widetilde{O}(n^{2^{-p}})$.

The only lower bound for random order comes from [GM07], where it is shown that a one pass algorithm requires space $\widetilde{\Omega}(\sqrt{n})$. The nature of this proof prevents it from generalizing to more than one pass, due to fundamental technical reasons; see the discussion on technical contributions below.

In this paper, we show the following tight lower bounds, addressing the 30-year-old question of Munro and Paterson:

THEOREM 1.1. *Consider a stream of n numbers, chosen from $[2n]$ uniformly at random, without replacement, and ordered uniformly at random. If an algorithm uses memory $O(n^{2^{-p}})$ and can, with error probability $1/3$, approximate the median by outputting a value with rank $\frac{n}{2} \pm n^{2^{-p}}$, then the algorithm must use $\Omega(p)$ passes. In particular, finding the median with $\text{polylog}(n)$ space requires $\Omega(\log \log n)$ passes.*

1.1 The Random-Order, Multipass Model.

Historically, the model that we are considering has found a place in both the theoretical and practical worlds; see the seminal paper of Munro and Paterson [MP80], as well as the very thorough justification in [GM06]. However, we find it worthwhile to reiterate why the model is appealing.

Streaming algorithms are designed with two classes of applications in mind: scenarios where data passes by and one cannot remember it all (e.g. in a network router), and scenarios where the data is on secondary storage and a memory-limited algorithm examines it sequentially. The second application allows multiple passes and is common

*Part of this work was done while the authors were visiting IBM Almaden Research Center. The first author is supported in part by NSF CAREER Award CCF-0448277 and a Dartmouth College Junior Faculty Fellowship.

[1] We let $\widetilde{O}(\cdot)$ ignore $\text{polylog}(n)$ factors. For simplicity, we assume the universe $[U] = \{1, \ldots, U\}$ for the numbers in the stream satisfies $U = \text{poly}(n)$, so that items in the stream can be stored in $O(\log n)$ bits.

in very large databases; this is the case where our lower bound is relevant.

While streaming is the only possible model for a router, any doubts as to whether it is a model worth studying for large databases have been erased by recent programming platforms like Google's Map-Reduce [DG04]. Forced by the massive, hugely distributed nature of data, these platforms essentially define streaming as the primitive for accessing data.

We now switch our attention to the random-order assumption of our model. Though worst-case order is the model of choice for theory, we believe the random-order model is an important link to the practical world which cannot be marginalized. The following are common cases in which random order is realized. While pondering these cases, the reader may want to consider a realistic example of a database query like "find quantiles for the salary of people hired between 2001 and 2004, and with the work location being India".

- random by assumption: if the values in the data set are assumed to come from some random source (e.g. salaries obey a distribution), the stream will have uniform ordering. This is an instance of average-case analysis, which is quite common in the realm of databases.
- random by heuristic: if the records in the database are ordered by some other keys (say, last name), then the order in which we read the interesting values (salary) is sufficiently arbitrary to be assumed random. This assumption is well-known in databases, as it is usually made by query optimizers.
- random by design: if we do not know enough about the query to build a useful data structure, we can decide to store records in a random order, hoping to avert worst-case behavior. This is the common playground for theory, which assumes worst-case data, but looks at the expected behavior over the algorithm's randomness.

The perspective of lower bounds. We believe the value of the random-order model for lower bounds is strictly higher that for upper bounds. A rather common objection to lower bounds in adversarial models are claims that they are irrelevant in real life, where heuristics do better since "data is never so bad in practice". Based on the real-world examples above, one may not be fully convinced that streams are truly random-ordered. However, the examples cast serious doubt that a very intricate worst-case order constructed by a lower bound is sufficiently plausible to mean anything in practice. The relevance of our lower bound is much harder to question, since it shows that even under the nicest assumptions about the data, an algorithm cannot do better.

1.2 Technical Contribution. For simplicity, define $\textsc{Stat}(i, S)$ to be the ith order statistic of the set S. Then, $\textsc{Median}(S) = \textsc{Stat}\big(\lfloor \frac{1}{2}|S| \rfloor, S\big)$. Also define $\textsc{Rank}(x, S) = |\{y \in S : y \le x\}|$, where it is not necessary that $x \in S$.

We first explain a very high-level intuition for the lower bound. When proving a lower bound for p passes, we break the stream into $p+1$ parts. We assign each part to a player and consider a communication game between these players. For example, we can say Player 1 receives the last $\sqrt{n}$ items, Player 2 the preceding $n^{7/8}$ items, Player 3 the preceding $n^{31/32}$ and so on, up to player $p+1$ who receives the remaining $\Omega(n)$ items at the beginning. Let T_i be the elements of player i, and $T_{\ge i} = \bigcup_{j \ge i} T_j$.

Assume the median occurs in T_{p+1}, i.e. the first part of the stream, which happens with constant probability. Then, let $r_i = \textsc{Rank}(\textsc{Median}(T), T_{\ge i+1})$. After learning r_i, players $i+1, \dots, p+1$ actually want to solve the r_i order statistic problem on their common input $T_{\ge i+1}$.

The hardness of the problem comes from the fact that r_i depends quite heavily on T_i. For example, if Player 1 sees the last $\sqrt{n}$ elements of the stream, an expected $\sqrt{n}/2$ of these elements are below the median. However, the number of elements below is actually a binomially distributed random variable with standard deviation $O(n^{1/4})$. Then, Player 1's input can shift r_1 anywhere in a $O(n^{1/4})$ range with roughly uniform probability. In this uncertainty range, Player 2 has an expected $|T_2|/\Theta(n^{3/4}) = n^{1/8}$ elements. Even when r_1 is known, these elements shift r_2 around by a binomial with standard deviation $\Theta(n^{1/16})$, etc.

The intuition for the hardness of tracing this sequence of $r_1, r_2, \dots,$ is a standard pointer chasing intuition. Remember that in each pass, the players speak in the order $p+1, p, \dots, 1$. If r_1 has an uncertainty range of $O(n^{1/4})$, it means players $p+1, \dots, 2$ must prepare to solve on the order of $n^{1/4}$ different order statistic problems, without having any idea which one until Player 1 speaks. Then, if they only communicate $\text{polylog}(n)$ bits each, they cannot in expectation gain significant knowledge about the order-statistic problem that turns out to be relevant. Then, in the second round we have the same intuition, with Player 2's input deciding r_2, and so on.

Problem structure. The hardness intuition presented above presumably dates back to Munro and Paterson, who conjectured that $\Theta(\log \log n)$ was the correct bound without having an upper bound. However, there are significant obstacles in the road to a lower bound, which accounts for why the conjecture has remained unresolved despite very significant progress in streaming lower bounds.

The challenges faced by the proof fall in two broad categories: understanding the structure of the problem correctly, and developing the right ideas in communication complexity to show that this structure is hard. With regard to the former, it should be noted that the intuition is shaky in many regards. For example, while it is true that players ≥ 2 are trying to compute $\text{STAT}(r_1, T_{\geq 2})$, the value of r_1 cannot be defined without reference to $\text{MEDIAN}(T)$. This defeats the purpose, since knowing r_1 may reveal significant information about the median.

We circumvent problems of this nature by identifying a more subtle and loose structure that makes the problem hard, while not deviating too far from the pointer chasing intuition. To demonstrate the obstacles that need to be surmounted, we mention that at a crucial point, our argument needs to invoke a nondeterministic prover that helps the players see the hard structure, while allowing the prover to communicate little enough to not make the problem easier.

Communication complexity. Still, the challenges regarding communication complexity are the more serious ones. At a high level, our proofs have the same flavor as the round elimination lemma [MNSW98, Sen03]. In the simplest incarnation of this lemma, Alice has a vector of B inputs $(x_1, \ldots, x_B)$, and Bob has an input y and an index $i \in [B]$. The players are trying to determine some $f(x_i, y)$. Alice speaks first, sending a message of $S \ll B$ bits. Then, this message is essentially worthless and can be eliminated, because it is communicating almost no information about the useful x_i, for random i.

Our situation is similar: r_i selects the next problem among $n^{\Omega(1)}$ choices, and players $\geq i+1$ don't know what to communicate that would be useful to the r_i problem. However, our setting is much more difficult because the random order of the stream forces a very particular distribution on the problem. Not only is the index r_i not uniform in its range (rather, it obeys a binomial distribution), but the problems x_i are heavily correlated. To see that, note for example the strong correlation between $\text{STAT}(r_i, T_{\geq i+1})$ and $\text{STAT}(r_i + 1, T_{\geq i+1})$.

The usual proofs of round elimination, based on manufacturing inputs to match a random message, cannot work here because of these correlations. At a technical level, note that round elimination normally imposes a nonproduct distribution on the problem, whereas we will need to have a product distribution.

By contrast, the only previous lower bound for our problem [GM07] was not based on understanding this unusual setting for round elimination, but on reducing a simple case of round elimination (indexing) to finding the median with one pass. While this is possible (with a lot of technical effort) for one pass, it fails entirely for multiple passes.

Our effort is part of a larger trend in recent literature. While the basic round elimination lemma is understood, many variants with peculiar requirements are studied, motivated by fundamental algorithmic questions for which we need lower bounds. For example, in FOCS 2004, Chakrabarti and Regev [CR04] study a variant tailored for approximate nearest neighbor; in SODA 2006, Adler et al. [ADHP06] study a variant tailored for distributed source coding and sensor networks; in STOC 2006 and SODA 2007 Pătraşcu and Thorup [PT06, PT07] study a variant tailored for predecessor search; Chakrabarti in CCC 2007 [Cha07] and Viola and Wigderson in FOCS 2007 [VW07] study variants tailored for multiparty number-on-the-forehead pointer chasing.

These papers, as well as the present one, push the theory of round elimination in very different directions, making it applicable to natural and interesting problems. Since round elimination has become a staple of modern communication complexity, one cannot help but compare this line of work to another well-developed area: the study of PCPs with different properties, giving inapproximability results for various problems.

2 Preliminaries

In this section, we boil down the task of proving our streaming lower bound to that of lower bounding the communication complexity of a related problem under a specific product distribution on its inputs.

2.1 Hard Instance.

DEFINITION 2.1. *The stream problem* RANDMEDIAN *is defined as follows. The input is a stream of n integers in $[2n]$ ordered uniformly at random. The desired output is any integer from the stream with rank between $\frac{n}{2} - \Delta$ and $\frac{n}{2} + \Delta$.*

DEFINITION 2.2. *For a permutation $\pi \in \mathcal{S}_{2n}$, we define the stream problem MED^π as follows. The input is a set $T \subset [2n]$ with $|T| = n$ that is presented as follows. Let $x \in \{0,1\}^{2n}$ be the characteristic vector of T. The input stream is $\langle x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(2n)} \rangle$, i.e., the bits of x ordered according to π. The desired output is* any *value $\mathcal{A}$ such that $|\text{RANK}(\mathcal{A}, T) - \frac{n}{2}| \leq \Delta$, i.e., a Δ-approximate median of T. A* random instance *of MED^π is defined to be one where T is chosen uniformly at random amongst all subsets of size n. Note that π continues to be fixed* a priori *and parametrizes the problem.*

We will construct a family $\mathcal{F} \subset \mathcal{S}_{2n}$ of permutations that consists of almost all of $\mathcal{S}_{2n}$ ($\mathcal{F}$ is to be thought of as a family of "typical permutations"). We shall then show a lower bound for MED^π for *any* $\pi \in \mathcal{F}$. Then, by contradiction with following lemma, we obtain a lower bound for RANDMEDIAN.

LEMMA 2.3. *Let $\mathcal{F} \subset \mathcal{S}_{2n}$ be a family of permutations with $|\mathcal{F}| \geq (1 - o(1)) \cdot (2n)!$. If* RANDMEDIAN *admits an ε-error p-pass streaming algorithm with space s, then* *there exists $\pi \in \mathcal{F}$ such that a random instance of* $\textsc{Med}^\pi$ *admits a p-pass streaming algorithm with distributional error $\varepsilon + o(1)$ and space $s + O(\log n)$.*

Proof. Let $\mathcal{A}$ be the algorithm for $\textsc{RandMedian}$ in the hypothesis. We propose the following algorithm $\mathcal{B}^\pi$ for $\textsc{Med}^\pi$: the input stream of bits $\langle x_{\pi(1)}, \ldots, x_{\pi(2n)} \rangle$ is transformed into the stream of integers $\langle \pi(i) : x_{\pi(i)} = 1 \rangle$, using $O(\log n)$ additional space, and fed to $\mathcal{A}$ as input. The output of $\mathcal{B}^\pi$ is the same as that of $\mathcal{A}$. Clearly, $\mathcal{B}^\pi$ is correct whenever $\mathcal{A}$ is. The key observation is that if π is distributed uniformly at random in $\mathcal{S}_{2n}$, then the input to $\mathcal{A}$ constructed by $\mathcal{B}^\pi$ is ordered uniformly at random. (This holds for every *fixed* $x \in \{0,1\}^{2n}$ and hence for a random x.) Therefore, the expected distributional error of $\mathcal{B}^\pi$ for such random π is at most ε. An averaging argument now shows that there exists $\pi \in \mathcal{F}$ such that the distributional error of $\mathcal{B}^\pi$ is at most $\varepsilon|\mathcal{S}_{2n}|/|\mathcal{F}| = \varepsilon + o(1)$. ∎

2.2 Communication Complexity. We now transform $\textsc{Med}^\pi$ into a multiparty number-in-hand communication game $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$ that is additionally parametrized by an integer $\widetilde{p}$, a vector $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_{\widetilde{p}}) \in \mathbb{N}^t$ and a vector $\boldsymbol{t} = (t_1, \ldots, t_{\widetilde{p}}) \in \mathbb{N}^t$. These parameters are required to satisfy:

$$\widetilde{p} \geq 2\,, \quad \sum_{i=1}^{\widetilde{p}} \ell_i = 2n\,, \quad \sum_{i=1}^{\widetilde{p}} t_i = n\,, \quad \text{and } \forall\, i \in [\widetilde{p}] : \binom{\ell_i}{t_i} \geq \frac{2^{\ell_i}}{n^2}\,. \tag{2.1}$$

Recall that the input to the stream problem is a bitvector $x \in \{0,1\}^{2n}$ ordered according to π. In the communication game, there are $\widetilde{p}$ players and Player i receives ℓ_i of the bits of x. Player 1 receives the last ℓ_1 bits of x according to π, Player 2 the ℓ_2 bits before that, etc. In other words, Player i receives the bits x_k for $k \in P_i^\pi$, where $P_1^\pi := \{\pi(n - \ell_1 + 1), \ldots, \pi(n)\}$, $P_2^\pi := \{\pi(n - \ell_1 - \ell_2 + 1), \ldots, \pi(n - \ell_1)\}$, etc.

The players communicate by writing s-bit messages on a blackboard, i.e. all messages are visible to all players. The games consists of p rounds. In each round, the players communicate in the fixed order $\widetilde{p}, (\widetilde{p}-1), \ldots, 2, 1$. The desired output is the same as that for $\textsc{Med}^\pi$ and must be written by Player 1 at the end of round p.

DEFINITION 2.4. *For an input $x \in \{0,1\}^{2n}$ to* $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$*, define $y_i \in \{0,1\}^{\ell_i}$ to be Player i's input, i.e., the projection of x on to the co-ordinates in P_i^π. Define the sets $T_i := \{k \in P_i^\pi : x_k = 1\}$, $T := \bigcup_i T_i$ and $T_{\geq i} := \bigcup_{j \geq i} T_j$.*

DEFINITION 2.5. *A* random instance *of* $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$ *is one where, for each i, y_i is chosen uniformly at random from the set $X_i := \{y \in \{0,1\}^{\ell_i} : |y| = t_i\}$.*

LEMMA 2.6. *Suppose $\boldsymbol{\ell}$ is such that $\ell_i = \Omega(\log n)$ for all i. If, for a particular π, a random instance of* $\textsc{Med}^\pi$ *has an ε-error p-pass streaming algorithm with space s, then there exists a suitable $\boldsymbol{t}$ satisfying conditions (2.1) such that a random instance of* $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$ *has an $(\varepsilon + o(1))$-error p-round communication protocol with message size s.*

Proof. A streaming algorithm for $\textsc{Med}^\pi$ translates in an elementary way into a communication protocol for $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$: the players simulate the streaming algorithm on their respective portions of the input, with each pass being simulated by one round of communication. They ensure continuity by communicating the memory contents of the streaming algorithm. This transformation incurs no additional error, but it only gives low *expected* distributional error for a suitably *random* $\boldsymbol{t}$. To be precise, suppose $x \sim \mathcal{U}'$, where $\mathcal{U}'$ denotes the uniform distribution on weight-n bitvectors in $\{0,1\}^{2n}$. Let y_i be as in Definition 2.4. Define the random variables $\hat{t}_i(x) = |y_i|$ and the random vector $\hat{\boldsymbol{t}}(x) = (\hat{t}_1(x), \ldots, \hat{t}_{\widetilde{p}}(x))$. For a vector $\boldsymbol{t}$ in the support of $\hat{\boldsymbol{t}}(x)$, let $\rho(\boldsymbol{t})$ be the distributional error of the above protocol for a random instance of $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$. Then, by the correctness guarantee of the streaming algorithm, we have $\mathbf{E}_{x \sim \mathcal{U}'}[\rho(\hat{\boldsymbol{t}}(x))] \leq \varepsilon$. (The reader may want to carefully compare the definition of a random instance of $\textsc{Med}^\pi$ with that of $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$.)

To prove the lemma, we must show the existence of a vector $\boldsymbol{t}$ such that $\boldsymbol{t}$ satisfies the conditions (2.1) and $\rho(\boldsymbol{t}) \leq \varepsilon + o(1)$. Let us call a particular $\boldsymbol{t}$ in the support of $\hat{\boldsymbol{t}}(x)$ *good* if it satisfies the final condition in (2.1) and *bad* otherwise. Let $\mathcal{U}$ be the uniform distribution on $\{0,1\}^{2n}$. Then,

$$\Pr_{x \sim \mathcal{U}'}[\hat{\boldsymbol{t}}(x) \text{ is bad}] = \Pr_{z \sim \mathcal{U}}[\hat{\boldsymbol{t}}(z) \text{ is bad} \mid |z| = n] \leq \frac{\Pr_{z \sim \mathcal{U}}[\hat{\boldsymbol{t}}(z) \text{ is bad}]}{\Pr_{z \sim \mathcal{U}}[|z| = n]} = \Pr_{z \sim \mathcal{U}}[\hat{\boldsymbol{t}}(z) \text{ is bad}] \cdot O(\sqrt{n})\,.$$

For $i \in [\widetilde{p}]$ and $z \sim \mathcal{U}$, $\hat{t}_i(z)$ is the weight of a uniformly random bitvector in $\{0,1\}^{\ell_i}$. Define the sets $I_i := \left\{t \in \{0,1,\ldots,\ell_i\} : \binom{\ell_i}{t} < \frac{2^{\ell_i}}{n^2}\right\}$. Then, by a union bound,

$$\Pr_{z \sim \mathcal{U}}[\hat{\boldsymbol{t}}(z) \text{ is bad}] \leq \sum_{i=1}^{\widetilde{p}} \sum_{t \in I_i} \Pr[\hat{t}_i(z) = t] = \sum_{i=1}^{\widetilde{p}} \sum_{t \in I_i} 2^{-\ell_i} \binom{\ell_i}{t} \leq \sum_{i=1}^{\widetilde{p}} \frac{\ell_i}{n^2} = \frac{2}{n}\,.$$

So,
$$\mathop{\mathbf{E}}_{x \sim \mathcal{U}'}[\rho(\hat{\boldsymbol{t}}(x)) \mid \hat{\boldsymbol{t}}(x) \text{ is good}] \leq \frac{\mathbf{E}_{x \sim \mathcal{U}'}[\rho(\hat{\boldsymbol{t}}(x))]}{\Pr_{x \sim \mathcal{U}'}[\hat{\boldsymbol{t}}(x) \text{ is good}]} \leq \frac{\varepsilon}{1 - \frac{1}{n} \cdot O(\sqrt{n})} = \varepsilon + o(1)\,.$$

Thus, there exists a good $\boldsymbol{t}$ such that $\rho(\boldsymbol{t}) \leq \varepsilon + o(1)$. ∎

2.3 The Permutation Family. We are now ready to define the permutation family $\mathcal{F}$ for which we prove the lower bound. Informally, $\mathcal{F}$ is the set of all permutations π for which each P_i^π is rather uniformly distributed in the range $[2n]$. Specifically, we break the range $[2n]$ into ℓ_i equal-sized buckets and insist that any k consecutive buckets contain $\Theta(k)$ elements of P_i^π for $k = \Omega(\log n)$. Formally:

$$\mathcal{F} = \Big\{\pi \in \mathcal{S}_{2n} : \forall\, i \in [\widetilde{p}], k = \Omega(\log n), j \le \ell_i - k, \\ \text{we have } \frac{k}{2} \le \Big|P_i^\pi \cap \Big[j \cdot \frac{n}{\ell_i}, (j+k)\cdot\frac{n}{\ell_i}\Big]\Big| \le 2k\Big\}$$

LEMMA 2.7. *Suppose $\ell_i = \Omega(\log n)$ for all i. Then* $\Pr_{\pi\in\mathcal{S}_n}[\pi \in \mathcal{F}] = 1 - o(1)$.

Proof. Pick $\pi \in \mathcal{S}_n$ uniformly at random, so that P_i^π is a random subset of $[n]$ of size ℓ_i. Let $A_{ijk} = P_i^\pi \cap [jn/\ell_i, (j+k)n/\ell_i]$. Clearly, for all (i,j,k), we have $\mathbf{E}[|A_{ijk}|] = k$. Applying a Chernoff-Hoeffding bound for the hypergeometric distribution, we have

$$\alpha_{ijk} := \Pr[|A_{ijk}| \notin [k/2, 2k]] < e^{-\Omega(k)} \le n^{-4},$$

where the latter inequality holds for $k = \Omega(\log n)$. A union bound over all $O(n^2\widetilde{p}) = O(n^3)$ triples (i,j,k) shows that $\Pr[\pi \notin \mathcal{F}] \le \sum_i\sum_j\sum_k \alpha_{ijk} = o(1)$. ∎

3 Warm-up: One Pass

In this section, we show that a one-pass algorithm for RANDMEDIAN either uses space $\Omega(n^{1/12})$, or requires approximation $\Delta = \Omega(n^{1/12})$. While this result is weaker than the previous lower bound for one pass [GM07], it demonstrates the basic structure of our argument in a simple case, and introduces some lemmas that will be required later.

By Lemmas 2.3 and 2.6, and Yao's minimax principle [Yao77], it suffices to show the following for some choice of $\widetilde{p}$ and $\boldsymbol{\ell}$: for all $\pi \in \mathcal{F}$ and $\boldsymbol{t} \in \mathbb{N}^{\widetilde{p}}$ satisfying condition (2.1), a $\frac{1}{3}$-error 1-round deterministic communication protocol for $\text{MEDCOMM}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$ with message size s must have $s = \Omega(n^{1/12})$. For the rest of this section, let us fix such a protocol. We also fix $\widetilde{p} = 2$. Let T and T_i be as in Definition 2.4 and chosen at random as in Definition 2.5. Let $\mathcal{A}$ be the random variable indicating the output of the protocol (which is an element of T).

Here is an outline of our proof. The protocol's guarantee is that $\Pr[|\text{RANK}(\mathcal{A}, T) - \frac{n}{2}| > \Delta] \le \frac{1}{3}$. We first fix Player 2's message, thereby restricting T_2 within some large subset $\mathcal{X}_2 \subset X_2$, and adding $o(1)$ error (Lemma 3.2). At this point $\mathcal{A}$ is a function of T_1 alone. Next, we define a quantity $r_1(T_1)$ that estimates $\text{RANK}(\mathcal{A}, T_2)$ to within about $\pm\Delta$, provided s is small (Corollary 3.4) and that takes on a large number $\Omega(\sqrt{\ell_1})$ of distinct values as we vary T_1 (Lemma 3.5). It then follows that we can find a large number, $B = \Omega(\sqrt{\ell_1}/\Delta^2)$, of instantiations of T_1 such that the corresponding r_1 values are $\Omega(\Delta^2)$ apart. Using the estimator property of $r_1(T_1)$, we then show that the corresponding values of $\mathcal{A}$ must also be $\Omega(\Delta^2)$ apart. This gap is large enough that if $B \gg s$, the corresponding random variables $\text{RANK}(\mathcal{A}, T_2)$ are "nearly independent" and have sufficient variance that they are unlikely to be confined within intervals of length at most 2Δ; the precise version of this statement is a key probabilistic fact that we call the Dispersed Ranks Theorem (Theorem 4.1). However, by the correctness guarantees, it *is* quite likely that the values of $\text{RANK}(\mathcal{A}, T_2)$ are so confined. This contradiction shows that $B = O(s)$, yielding the desired lower bound.

We now fill in the details, starting with the precise definition of r_1.

DEFINITION 3.1. *For $S \subset [2n]$, define $r_1(S) := \frac{n}{2} - |\{x \in S : x \le n\}| = \frac{n}{2} - \text{RANK}(n, S)$. Note that $n = \mathbf{E}[\text{MEDIAN}(T)]$.*

LEMMA 3.2. *If $s \ge \log n$, there exists $\mathcal{X}_2 \subset X_2$ such that the message sent by Player 2 is constant over $\mathcal{X}_2$, $|\mathcal{X}_2| \ge |X_2|/2^{2s}$, and $\Pr[|\text{RANK}(\mathcal{A}, T) - \frac{n}{2}| > \Delta \mid T_2 \in \mathcal{X}_2] \le \frac{1}{3} + \frac{1}{n}$.*

Proof. Since Player 2's message is s bits long, it partitions X_2 into 2^s subsets $\mathcal{X}_2^{(1)}, \ldots, \mathcal{X}_2^{(2^s)}$ such that the message is constant on each $\mathcal{X}_2^{(i)}$. Define

$$p_i := \Pr\Big[\Big|\text{RANK}(\mathcal{A}, T) - \frac{n}{2}\Big| > \Delta \mid T_2 \in \mathcal{X}_2^{(i)}\Big].$$

The protocol's guarantee implies $\sum_{i=1}^{2^s} p_i|\mathcal{X}_2^{(i)}|/|X_2| \le \frac{1}{3}$. Call an integer $i \in [2^s]$ *good* if $p_i \le \frac{1}{3} + \frac{1}{n}$ and *bad* otherwise. By Markov's inequality,

$$\sum_{i \text{ bad}} \frac{|\mathcal{X}_2^{(i)}|}{|X_2|} \le \frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{n}} = 1 - \frac{3}{n+3} \le 1 - \frac{1}{n};$$

$$\text{whence } \sum_{i \text{ good}} \frac{|\mathcal{X}_2^{(i)}|}{|X_2|} \ge \frac{1}{n}.$$

Therefore, there exists a good i such that $|\mathcal{X}_2^{(i)}|/|X_2| \ge 2^{-s}\cdot\frac{1}{n} \ge 4^{-s}$, where we used $s \ge \log n$. Setting $\mathcal{X}_2$ to be this particular $\mathcal{X}_2^{(i)}$ completes the proof. ∎

For the rest of this section, we fix an $\mathcal{X}_2$ with the properties guaranteed by the above lemma.

LEMMA 3.3. *If $s \ge \log n$ and $\ell_1 \le \sqrt{n}$, then* $\Pr\big[|\text{RANK}(\text{MEDIAN}(T), T_2) - r_1(T_1)| > 10s \mid T_2 \in \mathcal{X}_2\big] \le \frac{1}{n}$.

Proof. Note that $|\textsc{Rank}(\textsc{Median}(T), T_2) - r_1(T_1)| = |\textsc{Rank}(\textsc{Median}(T), T_2) + \textsc{Rank}(n, T_1) - \frac{n}{2}| = |\textsc{Rank}(n, T_1) - \textsc{Rank}(\textsc{Median}(T), T_1)| = |T_1 \cap I|$, where I is the interval between n and $\textsc{Median}(T)$. We now study the quantity $p_\lambda := \Pr[|n - \textsc{Median}(T)| > \lambda\sqrt{n} \mid T_2 \in \mathcal{X}_2]$, where $\lambda > 0$ is a real parameter.

To this end, let U be a uniform random subset of $[2n]$. Then, a simple Chernoff bound shows that $\Pr[|n - \textsc{Median}(U)| > \lambda\sqrt{n}] \leq e^{-\lambda^2/20}$. Note that the distribution of T is the same as that of U conditioned on the event $\mathcal{E}$:= "$\forall\, i \in \{1,2\} : |U \cap P_i^\pi| = t_i$." Clearly $\Pr[\mathcal{E}] = \prod_{i=1}^{2} 2^{-\ell_i}\binom{\ell_i}{t_i} \geq n^{-4}$, where we used the fact that the vector (t_1, t_2) satisfies condition (2.1). Furthermore, $\Pr[T_2 \in \mathcal{X}_2] \geq 2^{-2s}$, by Lemma 3.2. Therefore

$$p_\lambda = \Pr\left[|n - \textsc{Median}(U)| > \lambda\sqrt{n} \mid \mathcal{E}, T_2 \in \mathcal{X}_2\right] \leq \frac{e^{-\lambda^2/20}}{n^{-4}2^{-2s}}.$$

Now, set $\lambda = 10\sqrt{s}$. This gives $p_\lambda \leq n^4 2^{2s} e^{-5s} \leq n^4 2^{-5s} \leq \frac{1}{n}$, for $s \geq \log n$. Therefore, except with probability $\frac{1}{n}$, the interval I has length at most $10\sqrt{sn}$. If we break the range $[2n]$ into ℓ_1 equal-sized buckets, the interval I will fit into a union of at most $10\sqrt{sn}\cdot\ell_1/(2n) \leq 5\sqrt{s} \leq 5s$ consecutive buckets. By the defining property of $\mathcal{F}$, this means $|P_1^\pi \cap I| \leq 10s$. Since $T_1 \subset P_1^\pi$, we are done. ∎

COROLLARY 3.4. $\Pr\left[|\textsc{Rank}(\mathcal{A}, T_2) - r_1(T_1)| > \Delta + 10s \mid T_2 \in \mathcal{X}_2\right] \leq \frac{1}{3} + o(1)$.

Proof. Suppose $|\textsc{Rank}(\mathcal{A}, T) - \frac{n}{2}| \leq \Delta$ and $|\textsc{Rank}(\textsc{Median}(T), T_2) - r_1(T_1)| \leq 10s$. By Lemmas 3.2 and 3.3, it suffices to show that these conditions imply $|\textsc{Rank}(\mathcal{A}, T_2) - r_1(T_1)| \leq \Delta + 10s$. To do so, note that the former condition is saying that there are at most Δ values in T between $\mathcal{A}$ and $\textsc{Median}(T)$. Since $T_2 \subset T$, we have $|\textsc{Rank}(\mathcal{A}, T_2) - \textsc{Rank}(\textsc{Median}(T), T_2)| \leq \Delta$. Now we simply apply a triangle inequality. ∎

The above lemma establishes the "estimator property" of $r_1(T_1)$ mentioned earlier. We now show that r_1 has high variability even when restricted to inputs T_1 on which the protocol does not err much. For sets $S \subset [2n]$, define

$$e(S) := \Pr\left[|\textsc{Rank}(\mathcal{A}, T_2) - r_1(T_1)| > \Delta + 10s \mid T_2 \in \mathcal{X}_2, T_1 = S\right],$$

i.e, the error probability of the protocol when Player 1's input is S and Player 2 sends his fixed message corresponding to $\mathcal{X}_2$.

LEMMA 3.5. *Define* $R := \{r_1(T_1) : e(T_1) \leq 0.35\}$. *Then* $|R| = \Omega(\sqrt{\ell_1})$.

Proof. By Corollary 3.4, $\mathbf{E}_{T_1}[e(T_1)] \leq \frac{1}{3} + o(1)$. So, by a Markov bound, $\Pr_{T_1}[r_1(T_1) \in R] \geq 0.01$. To obtain the conclusion, we now show that $\Pr[r_1(T_1) \in R] = O(|R|/\sqrt{\ell_1})$.

As before, it is sufficient to perform an analysis under the assumption that $T_1 \subset P_1^\pi$ is chosen by including every element with probability $1/2$, independently. Examining Definition 3.1, we see that

$$r_1(T_1) = \frac{n}{2} - \textsc{Rank}(n, T_1) = \frac{n}{2} - |P_1^\pi \cap [n]|.$$

By the defining property of $\mathcal{F}$, we have $|P_1^\pi \cap [n]| = \Theta(\ell_1)$, so r_1 has a binomial distribution with standard deviation $\Theta(\sqrt{\ell_1})$. It follows that, for any x, $\Pr[r_1(T_1) = x] \leq O(1/\sqrt{\ell_1})$. Therefore $\Pr[r_1(T_1) \in R] = O(|R|/\sqrt{\ell_1})$, as desired. ∎

We can now prove the one-pass lower bound as follows. Define $\Delta_1 := \Delta + 10s$. For any $N \geq 1$, we can clearly find $|R|/N$ elements in R such that any two are at least N apart. Combining this simple observation with Lemma 3.5, we see that there exist instantiations $T_1^{(1)}, \ldots, T_1^{(B)}$ of the random set T_1, with $B \geq |R|/(100\Delta_1^2) = \Omega(\sqrt{\ell_1}/\Delta_1^2)$, such that

1. for all $i \in [B]$, $e\big(T_1^{(i)}\big) \leq 0.35$, and
2. for all $i \in [B-1]$, $r_1\big(T_1^{(i+1)}\big) - r_1\big(T_1^{(i)}\big) \geq 100\Delta_1^2$ and $r_1\big(T_1^{(1)}\big) \geq 100\Delta_1^2$.

Recall that $\mathcal{X}_2$ has been fixed, so $\mathcal{A}$ is a function of T_1 alone. Let $\mathcal{A}^{(i)}$ be the output of the protocol when $T_1 = T_1^{(i)}$ and, for convenience, define $\mathcal{A}^{(0)} = 0$. Define the intervals $R_i := \big[r_1\big(T_1^{(i)}\big) - \Delta_1, r_1\big(T_1^{(i)}\big) + \Delta_1\big]$. Pick any $i \in [B-1]$. By condition 1 above,

$$\Pr\left[\textsc{Rank}\big(\mathcal{A}^{(i)}, T_2\big) \in R_i \bigwedge \textsc{Rank}\big(\mathcal{A}^{(i+1)}, T_2\big) \in R_{i+1}\right] \geq 1 - e\big(T_1^{(i)}\big) - e\big(T_1^{(i+1)}\big) > 0.$$

Therefore, by condition 2, there exists an instantiation T_2^* of T_2 such that $\textsc{Rank}\big(\mathcal{A}^{(i+1)}, T_2^*\big) - \textsc{Rank}\big(\mathcal{A}^{(i)}, T_2^*\big) \geq 100\Delta_1^2 - 2\Delta_1 \geq 99\Delta_1^2$. Thus, $\mathcal{A}^{(i+1)} - \mathcal{A}^{(i)} \geq 99\Delta_1^2$. Similar reasoning shows that this inequality in fact holds for $i = 0$ as well. In summary, the values $\mathcal{A}^{(0)}, \ldots, \mathcal{A}^{(i)}$ are seen to be *well dispersed*.

On the other hand, condition 1 above can be written as

$$\forall\, i \in [B] : \Pr_U\left[\textsc{Rank}\big(\mathcal{A}^{(i)}, U\big) \in R_i \;\big|\; U \in \mathcal{X}_2\right] \geq 0.65,$$

where U denotes a uniform random subset of $[\ell_2]$. Let $\mathcal{E}^*$ denote the event $\big|\big\{i \in [B] : \textsc{Rank}\big(\mathcal{A}^{(i)}, U\big) \in R_i\big\}\big| \geq 0.6B$. A Markov bound gives us $\Pr[\mathcal{E}^* \mid U \in \mathcal{X}_2] \geq 1/8$. Furthermore,

$$\Pr[U \in \mathcal{X}_2] = 2^{-\ell_2}|\mathcal{X}_2| \geq 2^{-\ell_2 - 2s}|X_2| = 2^{-\ell_2 - 2s}\binom{\ell_2}{t_2} \geq \frac{2^{-2s}}{n^2},$$

where the final inequality used (2.1). Therefore, $\Pr[\mathcal{E}^*] \geq (1/8) \cdot 2^{-2s} n^{-2}$.

At this point we invoke a key probabilistic fact — the Dispersed Ranks Lemma — which says that for such well dispersed $\mathcal{A}^{(i)}$ values, we must have $\Pr[\mathcal{E}^*] \leq 2^{-\Omega(B)}$. Combined with the above lower bound on $\Pr[\mathcal{E}^*]$, this implies $s \geq \Omega(B) - O(\log n) = \Omega(\sqrt{\ell_1}/(\Delta + 10s)^2) - O(\log n)$. Setting $\ell_1 = \sqrt{n}$ (the maximum allowed by Lemma 3.3) and rearranging gives $\max\{s, \Delta\} = \Omega(n^{1/12})$, the desired lower bound.

4 The Dispersed Ranks Lemma

We now introduce a key technical probabilistic fact that lies at the heart of our lower bound argument and captures the intuition behind round elimination in our setting. The theorem was used in the above proof of the lower bound for one-pass algorithms. Later, it will be used repeatedly for the multipass lower bound.

LEMMA 4.1. (DISPERSED RANKS LEMMA) *Let ℓ and B be large enough integers and let U denote a uniform random subset of $[\ell]$. Let $q_0 = 0$ and let $q_1, q_2, \ldots, q_B \in [\ell]$ be such that $\forall\, i : q_{i+1} - q_i \geq 99\Delta^2$. Let $R_i := [q_i - \Delta, q_i + \Delta]$ for $i \in [B]$, and let $\mathcal{E}^*$ denote the event $\left|\{i \in [B] : \textsc{Rank}(q_i, U) \in R_i\}\right| \geq 0.6B$. Then $\Pr[\mathcal{E}^*] = 2^{-cB}$, for some constant $c > 0$.*

Proof. Let Z_i denote the random variable $\textsc{Rank}(q_i, U)$ for all $i \in [B]$. By the union bound,

$$\Pr[\mathcal{E}^*] \;\leq\; \sum_S \Pr\Big[\bigwedge_{i \in S} (Z_i \in R_i)\Big],$$

where S ranges over all subsets of $[B]$ containing exactly $0.6B$ indices. We will show that each probability within the sum is at most $2^{-c'B}$ for some constant $c' > 1$. Since the number of choices of S is at most 2^B, the proof of the lemma follows.

Fix a set $S \subseteq [B]$ of size $0.6B$. For each $i \in S$, define $J_i = \{j \in S \mid j < i\}$. By the chain rule of probability,

$$\begin{aligned} &\Pr\Big[\bigwedge_{i \in S} (Z_i \in R_i)\Big] \\ &\quad = \prod_{i \in S} \Pr\Big[Z_i \in R_i \;\Big|\; \bigwedge_{j \in J_i} (Z_j \in R_j)\Big] \end{aligned} \tag{4.2}$$

Fix an i in the above product in (4.2) above. Also fix a set of elements $z_j \in R_j$ for all $j \in J_i$. Let $\mathcal{E}$ denote the event $\bigwedge_{j \in J_i} (Z_j = z_j)$. We will upper bound the probability $\Pr[Z_i \in R_i \mid \mathcal{E}]$. By averaging, this will also yield the same upper bound on the probability in (4.2).

Let k denote the largest element in J_i. Conditioned on the event $\mathcal{E}$, Z_i is the sum of z_k and a binomially distributed random variable corresponding to a sum of $q_i - q_k$ independent Bernoulli random variables. By the well-separated property of the q_j's we have $q_i - q_k \geq 99\Delta^2$. Using the property of the binomial distribution, the probability that Z_i attains any value is at most $1/\sqrt{99\Delta^2}$. Therefore, $\Pr[Z_i \in R_i \mid \mathcal{E}] \leq |R_i|/\sqrt{99\Delta^2} \leq 2/\sqrt{99}$. Using this bound in (4.2), it follows that

$$\Pr\Big[\bigwedge_{i \in S} (Z_i \in R_i)\Big] \;\leq\; (2/\sqrt{99})^{|S|} \;=\; 2^{-c'B},$$

where $c' > 1$. ∎

5 Two Passes

In this section, we show that a 2-pass algorithm requires $\max\{s, \Delta\} = \Omega(n^{3/80})$. This proof contains all the ideas needed for the general lower bound for p passes. However, in this extended abstract, we choose to present the lower bound for $p = 2$, which allows for much more transparent notation and discussion. A proof of the general lower bound is deferred to the full version of the paper.

We fix the number of players $\widetilde{p} = 3$, for the communication problem $\textsc{MedComm}^{\pi,\boldsymbol{\ell},\boldsymbol{t}}$. In general, for a p-pass algorithm, we would fix $\widetilde{p} = p + 1$. Assume we have a $\frac{1}{3}$-error 2-round deterministic protocol for the problem with message size s.

We begin by fixing the first round of communication in essentially the same way as in the one-pass lower bound. First, we fix the messages of Players 3 and 2 as in Lemma 3.2. Now define $r_1(T_1)$ as before, and conclude that there exist $\Omega(\sqrt{\ell_1})$ settings of T_1, leading to distinct r_1 values, where the protocol's error is at most 0.35. To maximize hardness, pick B choices $r_1^1, \ldots, r_1^B$ for r_1 that are as far away as possible, i.e. for all k, $r_1^{k+1} - r_1^k = \Omega(\sqrt{\ell_1}/B)$.

We now consider B simulations for the second round, depending on the B picked choices of T_1 (more precisely, depending on the messages output by Player 1 given the B choices of T_1). Let $\mathcal{A}^1, \ldots, \mathcal{A}^B$ be the algorithm's output in all these simulations. Now, Player 3 sends B messages of s bits, effectively a Bs-bit message, which we fix as in Lemma 3.2. At the end of all these steps, we have:

$$\begin{aligned} (T_3, T_2) \in \mathcal{X}_3 \times \mathcal{X}_2,\; |\mathcal{X}_3| \geq |X_3|/2^{O(Bs)}, \\ |\mathcal{X}_2| \geq |X_2|/2^{O(s)}\,; \end{aligned} \tag{5.3}$$

$$\begin{aligned} \forall\, i : \Pr\Big[\left|\textsc{Rank}(\mathcal{A}^i, T_{\geq 2}) - r_1^i\right| \geq \Delta + O(s) \\ \mid (T_3, T_2) \in \mathcal{X}_3 \times \mathcal{X}_2\Big] \;\leq\; 0.35 + o(1)\,. \end{aligned} \tag{5.4}$$

As before, to show that the information about T_3 is not enough, we must analyze what problem is being solved from Player 3's perspective. That is, we want to understand $\textsc{Rank}(\mathcal{A}^i, T_3)$. By (5.4), we must understand $\textsc{Rank}(\textsc{Stat}(r_1^i, T_{\geq 2}), T_3)$. Let us define

$$\xi^i := \textsc{Stat}(r_1^i, T_{\geq 2})\,. \tag{5.5}$$

Intuitively speaking, r_1^i and r_1^{i-1} are separated by $\Omega(\sqrt{\ell_1}/B)$, so there are on the order of $\frac{\ell_2}{n} \cdot \frac{\sqrt{\ell_1}}{B}$ elements in P_2^π between ξ^i and ξ^{i-1}. This makes for a variance of $\textsc{Rank}(\xi^i, T_3)$ of roughly $\left(\frac{\ell_2}{n} \cdot \frac{\sqrt{\ell_1}}{B}\right)^{1/2}$. Since the variance needs to be high to make for a hard problem, we have imposed a lower bound for ℓ_2/n.

On the other hand, we need to show $\textsc{Rank}(\xi^i, T_3)$ has *small* variance conditioned on T_2. That is done by constructing a good estimator r_2 based on T_2. Our ability to do that depends on how well we can understand ξ^i. Specifically, if we understand it to within $\pm D$, we have $D\frac{\ell_2}{n}$ values in P_2 that cannot be compared reliably to ξ^i, so the estimator for $\textsc{Rank}(\xi^i, T_3)$ suffers an additive approximation of $D\frac{\ell_2}{n}$. To keep the approximation in check, we must impose an upper bound on ℓ_2/n.

Thus, we have forces upper bounding and lower bounding ℓ_2/n, and we must make sure that a good choice actually exists. That is done by constructing an estimator with small enough D. To make better estimation possible, we need some more information about the stream. It turns out that if we find out $\textsc{Median}(T_{\geq 2})$, we reduce the uncertainty range of ξ^i enough to get a good D. This is intuitive, since r_1^i is only $O(\sqrt{\ell_1})$ away from $\textsc{Median}(T_{\geq 2})$. However, obtaining $\textsc{Median}(T_{\geq 2})$ is hard in our model (that is essentially what we are trying to prove). To circumvent that, we note that it is an easy computation based on nondeterminism. On the other hand, a small intervention by a nondeterministic prover cannot help solve all r_1^i problems, so we still get a lower bound even if we allow nondeterminism in a brief part of the communication game.

5.1 Constructing an Estimator r_2. We now attempt to construct a good estimator $r_2(r_1^i, T_2)$ for the interesting quantity $\textsc{Rank}(\xi^i, T_3)$. In general, T_2 does not give enough information to construct a very good estimator r_2. However, we restrict the problem to a subset of the inputs where such an estimator exists. It turns out that the one critical piece of information that we need is $\textsc{Median}(T_2 \cup T_3)$, so we work in a set of the inputs where it is fixed.

LEMMA 5.1. *Let $\mathcal{X}_2 \subset X_2$, $\mathcal{X}_3 \subset X_3$ be arbitrary. There exist $\mathcal{X}_2' \subset \mathcal{X}_2, \mathcal{X}_3' \subset \mathcal{X}_3$ and a constant M such that:*

- $|\mathcal{X}_2'|/|\mathcal{X}_2| \geq 2^{-O(\log n)}$ *and* $|\mathcal{X}_3'|/|\mathcal{X}_3| \geq 2^{-O(\log n)}$*;*
- $\Pr\left[\left|\textsc{Rank}(\mathcal{A}^i, T_{\geq 2}) - r_1^i\right| \geq \Delta + O(s)\right] \leq 0.37$*;*
- $\textsc{Median}(T_2 \cup T_3) = M$ *for all* $(T_2, T_3) \in \mathcal{X}_2' \times \mathcal{X}_3'$.

Proof. Consider the following nondeterministic communication protocol for finding $\textsc{Median}(T_2 \cup T_3)$ with $O(\log n)$ communication. The prover proposes the median x and $\textsc{Rank}(x, T_2)$. Player 2 accepts iff this rank is correct. Player 3 accepts iff $\textsc{Rank}(x, T_3) + \textsc{Rank}(x, T_2) = (|T_2| + |T_3|)/2$.

Note that there is a unique acceptable witness (proof) for every problem instance. In other words, the nondeterministic protocol induces a *partition* of $\mathcal{X}_2 \times \mathcal{X}_3$ into a certain number, N_R, of rectangles. Let us discard all rectangles with size less than $|\mathcal{X}_2 \times \mathcal{X}_3|/(100N_R)$. At least a 0.99 fraction of the space $\mathcal{X}_2 \times \mathcal{X}_3$ survives, so the average error over this remainder of the space increases by at most 0.01. Pick any remaining rectangle over which the error increases by at most 0.01 and call it $\mathcal{X}_2' \times \mathcal{X}_3'$. Then, since $|\mathcal{X}_2' \times \mathcal{X}_3'| \geq |\mathcal{X}_2 \times \mathcal{X}_3|/(100N_R)$, we have $|\mathcal{X}_2'| \geq |\mathcal{X}_2|/(100N_R)$ and $|\mathcal{X}_3'| \geq |\mathcal{X}_3|/(100N_R)$. Finally, observe that $N_R = 2^{O(\log n)}$, since the nondeterministic protocol sends $O(\log n)$ bits. ∎

Henceforth, we shall fix the spaces $\mathcal{X}_2'$ and $\mathcal{X}_3'$ (and the constant M) guaranteed by the above lemma and work within them.

Assume by symmetry that $r_1^i \geq (t_2 + t_3)/2$, that is $\xi^i \geq M$. If we knew ξ^i, we could compute:

$$\begin{aligned}\textsc{Rank}(\xi^i, T_3) &= r_1^i - |\{y \in T_2 \mid y \leq \xi^i\}| \\ &= r_1^i - \textsc{Rank}(M, T_2) - \left|T_2 \cap [M, \xi^i]\right|.\end{aligned}$$

Since ξ^i is not known, we can proceed in the same way, using $\mathbf{E}[\xi^i]$ instead. Unfortunately, *a priori* ξ^i is not concentrated too tightly, and this uncertainty would introduce too large an approximation in the estimate of $|T_2 \cap [M, \xi^i]|$. However, this is precisely why we want to fix M: conditioned on $\textsc{Median}(T_{\geq 2}) = M$, ξ^i is much more tightly concentrated, and

$$\Xi^i := \mathbf{E}[\xi^i \mid \textsc{Median}(T_{\geq 2}) = M]$$

is a good enough replacement for the real ξ^i. We thus define:

$$\begin{aligned}r_2(r_1^i, T_2) &= r_1^i - \textsc{Rank}(\Xi^i, T_2) \\ &= r_1^i - \textsc{Rank}(M, T_2) - \left|T_2 \cap [M, \Xi^i]\right|.\end{aligned}$$

LEMMA 5.2. *For $\lambda \geq \Omega(B)$, we have:* $\Pr\left[\left|\xi^i - \Xi^i\right| \geq \lambda\sqrt[4]{\ell_1} \mid (T_3, T_2) \in \mathcal{X}_3' \times \mathcal{X}_2'\right] \leq 2^{-\Omega(\lambda^2)}$.

Proof. The following random walk defines $\xi_i = \textsc{Stat}(r_1^i, T_{\geq 2})$: start with $\textsc{Median}(T_{\geq 2})$ and go up on elements of $P_2^\pi \cup P_3^\pi$, until you find $r_1^i - \frac{t_2+t_3}{2}$ elements that are in $T_{\geq 2}$. The length of this walk is an approximate bound for $\xi^i - \textsc{Median}(T_{\geq 2})$. The only discrepancy is the number of elements in $[2n] \setminus (P_2^\pi \cup P_3^\pi) = P_1^\pi$ that are skipped. However, we will only be interested in walks whose length does not deviate too much from its expectation. Thus, the length is $O(r_1^i - \frac{t_2+t_3}{2}) \leq O(\sqrt{\ell_1})$. By the defining property of $\mathcal{F}$, there are only $O(\log n)$ elements of P_1^π in the relevant range, so the length of the walk is an $O(\log n)$ additive approximation to $\xi^i - \textsc{Median}(T_{\geq 2})$.

Now assume $T_{\geq 2}$ is selected from $P_2^\pi \cup P_3^\pi$ by including every element independently and uniformly. Then the length of the walk deviates from its expectation by $\lambda\sqrt{r_1^i - \frac{t_2+t_3}{2}} \leq \lambda\sqrt[4]{\ell_1}$ with probability $2^{-\Omega(\lambda^2)}$. The $O(\log n)$ approximation is a lower order term compared to $\lambda\sqrt[4]{\ell_1}$ (affecting constant factors in λ), so this is a bound on the deviation of $\xi^i - \textsc{Median}(T_{\geq 2})$ from its mean.

Now to obtain the real process of selecting $T_{\geq 2}$, all we have to do is condition on $|T_2| = t_2, |T_3| = t_3$. These events have probability $1/\operatorname{poly}(n)$ by (2.1), so the probability of the bad event is $\leq 2^{-\Omega(\lambda^2)} \cdot \operatorname{poly}(n)$. Since $\lambda = \Omega(B) > \log n$, we have $2^{-\Omega(\lambda^2)} \cdot \operatorname{poly}(n) = 2^{-\Omega(\lambda^2)}$. Now we condition on $(T_3, T_2) \in \mathcal{X}_3' \times \mathcal{X}_2'$. By (5.3) and Lemma 5.1, we have $\Pr[(T_3, T_2) \in \mathcal{X}_3' \times \mathcal{X}_2'] \geq 2^{-O(Bs)-O(\log n)} \geq 2^{-O(Bs)}$. So in the universe $\mathcal{X}_3' \times \mathcal{X}_2'$, the probability of a deviation is at most $2^{-\Omega(\lambda^2)}/2^{-O(Bs)}$. If $\lambda \geq \Omega(B)$, this is $2^{-\Omega(\lambda^2)}$.

Finally, note that in $\mathcal{X}_3' \times \mathcal{X}_2'$, $\textsc{Median}(T_{\geq 2})$ is fixed. Then, the event that $\xi^i - \textsc{Median}(T_{\geq 2})$ doesn't deviate from the expectation is the same as the event that ξ^i doesn't. ∎

COROLLARY 5.3. *If $\ell_2 < n/\sqrt[4]{\ell_1}$, then $\forall i \in [B]$, $\Pr\left[\left|\textsc{Rank}(\xi^i, T_3) - r_2(r_1^i, T_2)\right| > \Omega(B) \mid (T_3, T_2) \in \mathcal{X}_3' \times \mathcal{X}_2'\right] = o(1)$.*

Proof. Inspecting the definition of r_2, we observe that $\left|\textsc{Rank}(\xi^i, T_3) - r_2(r_1^i, T_2)\right| \leq \left|T_2 \cap [\xi^i, \Xi^i]\right|$. Setting $\lambda = \Theta(B)$ in Lemma 5.2, $|\xi^i - \Xi^i| \leq O(B) \cdot \sqrt[4]{\ell_1}$ entails $\left|T_2 \cap [\xi^i, \Xi^i]\right| \leq \left|P_2^\pi \cap [\xi^i, \Xi^i]\right| \leq (\ell_2/n) \cdot O(B\sqrt[4]{\ell_1}) = O(B)$, where the final inequality follows from the defining property of $\mathcal{F}$. ∎

5.2 The Variability of r_2. Putting the bounds on the estimator r_2 together, we obtain the following problem:

$$(T_3, T_2) \in \mathcal{X}_3 \times \mathcal{X}_2,\ |\mathcal{X}_3| \geq |X_3|/2^{O(Bs)},\quad |\mathcal{X}_2| \geq |X_2|/2^{O(s)}\,;$$

$$\forall i : \Pr\Big[\left|\textsc{Rank}(\mathcal{A}^i, T_3) - r_2(r_1^i, T_2)\right| \geq \Delta + O(B) \mid (T_3, T_2) \in \mathcal{X}_3 \times \mathcal{X}_2\Big] \ \leq\ 0.37 + o(1)\,.$$

Thus, we have identified the problem that the algorithm is solving with small approximation. The only remaining question is how many choices of r_2 exist, for possible choices of T_2. These are distinct problems that the messages from Player 3 must have answered.

We now sketch the remainder of the analysis, deferring a complete rigorous treatment to the full version of the paper. The analysis is completed by applying a version of the Dispersed Ranks Lemma to Player 2. Let G be the set of r_2 values that get generated by settings (T_2, T_1) which don't lead to error above 0.47. By Markov, $\Pr[r_2 \in G] \geq 1/10$. Note that $r_1^{k+1} - r_1^k = \Omega(\sqrt{\ell_1}/B)$ and there are $\Omega(\frac{\ell_2}{n} \cdot \sqrt{\ell_1}/B)$ values of P_2^π in this range. With $\ell_1 = \sqrt{n}$, $\ell_2 = n^{15/16}$, this gives $\Omega(n^{3/16}/B)$ values. So the lemma is applied with $\Delta = \Omega(n^{3/32}/\sqrt{B})$. The conclusion will be that an event of the form $\mathcal{E}^*$ is exponentially unlikely unless $|G| = \Omega(B\Delta) = \Omega(\sqrt{B} \cdot n^{3/32})$. Therefore, we have $\Omega(\sqrt{B} \cdot n^{3/32})$ possible values for r_2.

The proof is completed as before, using the dispersed ranks property for the last player. We need B^2 choices of r_2^i, so we can guarantee $r_2^{i+1} - r_2^i = \Omega(B\Delta/B^2) = \Omega(n^{3/32}/\sqrt{B})$. Then the new Δ for the lemma is $\Omega(n^{3/64}/\sqrt[4]{B})$, and we thus obtain an inapproximability of $n^{3/64}/\sqrt[4]{B}$. On the other hand, we have an upper bound for the approximation of $O(\Delta + B)$, so this is impossible if $\Delta = B$ and $B^{5/4} < n^{3/64}$, if $B < n^{3/80}$. That means $s + \Delta = \Omega(n^{3/80})$.

Acknowledgments

We are grateful to Sudipto Guha for suggesting to us the problem studied here, and for inspiring and motivating conversations in the early stages of this work.

References

[ADHP06] Micah Adler, Erik D. Demaine, Nicholas J. A. Harvey, and Mihai Pătraşcu. Lower bounds for asymmetric communication channels and distributed source coding. In *Proc. 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 251–260, 2006.

[Cha07] Amit Chakrabarti. Lower bounds for multi-player pointer jumping. In *Proc. 22nd Annual IEEE Conference on Computational Complexity*, pages 33–45, 2007.

[CKMS06] Graham Cormode, Flip Korn, S. Muthukrishnan, and Divesh Srivastava. Space- and time-efficient deterministic algorithms for biased quantiles over data streams. In *Proc. 25th ACM Symposium on Principles of Database Systems*, pages 263–272, 2006.

[CM05] Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and itsapplications. *J. Alg.*, 55(1):58–75, 2005. Preliminary version in *Proc. 6th Latin American Theoretical Informatics Symposium*, pages 29–38, 2004.

[CR04] Amit Chakrabarti and Oded Regev. An optimal randomised cell probe lower bound for approximate nearest neighbour searching. In *Proc. 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 473–482, 2004.

[DG04] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proc. 6th Symposium on Operating System Design and Implementation*, pages 137–150, 2004.

[GK01] Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. In *Proc. Annual ACM SIGMOD Conference*, pages 58–66, 2001.

[GKMS02] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. How to summarize the universe: Dynamic maintenance of quantiles. In *Proc. 28th International Conference on Very Large Data Bases*, pages 454–465, 2002.

[GM06] Sudipto Guha and Andrew McGregor. Approximate quantiles and the order of the stream. In *Proc. 25th ACM Symposium on Principles of Database Systems*, pages 273–279, 2006.

[GM07] Sudipto Guha and Andrew McGregor. Lower bounds for quantile estimation in random-order and multi-pass streaming. In *Proc. 34th International Colloquium on Automata, Languages and Programming*, pages 704–715, 2007.

[GZ03] Anupam Gupta and Francis Zane. Counting inversions in lists. In *Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 253–254, 2003.

[MNSW98] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998. Preliminary version in *Proc. 27th Annual ACM Symposium on the Theory of Computing*, pages 103–111, 1995.

[MP80] J. Ian Munro and Mike Paterson. Selection and sorting with limited storage. *TCS*, 12:315–323, 1980. Preliminary version in *Proc. 19th Annual IEEE Symposium on Foundations of Computer Science*, pages 253–258, 1978.

[MRL98] Gurmeet Singh Manku, Sridhar Rajagopalan, and Bruce G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *Proc. Annual ACM SIGMOD Conference*, pages 426–435, 1998.

[MRL99] Gurmeet Singh Manku, Sridhar Rajagopalan, and Bruce G. Lindsay. Random sampling techniques for space efficient online computation of order statistics of large datasets. In *Proc. Annual ACM SIGMOD Conference*, pages 251–262, 1999.

[PT06] Mihai Pǎtraşcu and Mikkel Thorup. Time-space trade-offs for predecessor search. In *Proc. 38th Annual ACM Symposium on the Theory of Computing*, pages 232–240, 2006.

[PT07] Mihai Pǎtraşcu and Mikkel Thorup. Randomization does not help searching predecessors. In *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 555–564, 2007.

[SBAS04] Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, and Subhash Suri. Medians and beyond: new aggregation techniques for sensor networks. In *Proc. ACM SenSys*, pages 239–249, 2004.

[Sen03] Pranab Sen. Lower bounds for predecessor searching in the cell probe model. In *Proc. 18th Annual IEEE Conference on Computational Complexity*, pages 73–83, 2003.

[VW07] Emanuele Viola and Avi Wigderson. One-way multiparty communication lower bound for pointer jumping with applications. In *Proc. 48th Annual IEEE Symposium on Foundations of Computer Science*, 2007. to appear.

[Yao77] Andrew C. Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proc. 18th Annual IEEE Symposium on Foundations of Computer Science*, pages 222–227, 1977.