High Entropy Random Selection Protocols[*]

Harry Buhrman [†] Matthias Christandl [‡] Michal Koucký [§] Zvi Lotker [¶]
Boaz Patt-Shamir [||] Nikolay Vereshchagin [**]

October 8, 2013

Abstract

We study the two party problem of randomly selecting a string among all the strings of length n. We want the protocol to have the property that the output distribution has high *entropy*, even when one of the two parties is dishonest and deviates from the protocol. We develop protocols that achieve high, close to n, entropy.

In the literature the randomness guarantee is usually expressed as being close to the uniform distribution or in terms of resiliency. The notion of entropy is not directly comparable to that of resiliency, but we establish a connection between the two that allows us to compare our protocols with the existing ones.

We construct an explicit protocol that yields entropy $n - O(1)$ and has $4\log^* n$ rounds, improving over the protocol of Goldreich et al. [5] that also achieves this entropy but needs $O(n)$ rounds. Both these protocols need $O(n^2)$ bits of communication.

Next we reduce the communication in our protocols. We show the existence, non-explicitly, of a protocol that has 6 rounds, $2n + 8\log n$ bits of communication and yields entropy $n - O(\log n)$ and min-entropy $n/2 - O(\log n)$. Our protocol achieves the same entropy bound as the recent, also non-explicit, protocol of Gradwohl et al. [6], however achieves much higher min-entropy: $n/2 - O(\log n)$ versus $O(\log n)$.

Finally we exhibit a very simple explicit protocol. We connect the security parameter of this protocol with the well studied Kakeya problem motivated by harmonic analysis and analytical number theory. We prove that this protocols has entropy $n - O(n^{3/4})$ and $n/2 - O(\log n)$ min-entropy. Therefore it performs almost as well with respect to the explicit constructions of Gradwohl et al. [6] entropy-wise, and has much better min-entropy. Its relation to the Kakeya problem follows a new and different approach to the random selection problem than any of the previously known protocols.

[*] The preliminary version of this paper was published in Proceedings of 10th International Workshop, APPROX 2007, and 11th International Workshop, RANDOM 2007, Princeton, NJ, USA, August 20-22, 2007. Proceedings. Lecture Notes in Computer Science, volume 4627/2007 pages 366-379.

[†] CWI and University of Amsterdam. Email: buhrman@cwi.nl.

[‡] Cambridge University. Email: mc380@cam.ac.uk, work done while visiting CWI.

[§] Mathematical Institute of Czech Academy of Sciences. Email: koucky@math.cas.cz, work done while visiting CWI.

[¶] Ben Gurion University, zvilo@cse.bgu.ac.il, work done while visiting CWI.

[||] Tel Aviv University. Email: boaz@eng.tau.ac.il

[**] Lomonosov Moscow State University. Email: ver@mech.math.msu.su, work done while visiting CWI.

1 Introduction

We study the following communication problem. Alice and Bob want to select a random string. They are not at the same location so they do not see what the other player does. They communicate messages according to some protocol and in the end they output a string of n bits which is a function of the messages communicated. This string should be as random as possible, in our case we measure the amount of randomness by the entropy of the probability distribution that is generated by this protocol.

The messages they communicate may depend on random experiments the players perform and on messages sent so far. The outcome of an experiment is known only to the party which performs it so the other party cannot verify the outcome of such an experiment or whether the experiment was carried out at all. One or both the parties may deviate from the protocol and try to influence the selected string (*cheat*). We are interested in the situation when a party honestly follows the protocol and wants to have some guarantee that the selected string is indeed as random as possible. The measure of randomness we use is the *entropy* of probability distribution that is the outcome of the protocol.

In this paper we present protocols for this problem. In particular we show a protocol that achieves entropy $n - O(1)$ if at least one party is honest and that uses $4 \log^* n$ rounds and communicates $n^2 + O(n \log n)$ bits. The round complexity of our protocol is optimal up-to a constant factor; the optimality follows from a result of Sanghvi and Vadhan [10]. We further consider the question of reducing the communication complexity of our protocol. We show non-constructively that there are protocols with linear communication complexity that achieve entropy $n - \log n$ in just 3 rounds, and in 6 rounds achieves in addition min-entropy $n/2 - O(\log n)$ which is close to the optimal bound of $n/2$, that follows from Goldreich et al. [5] and from a bound on quantum coin-flipping due to Kitaev (see [2]). We propose an explicit and very simple protocol that has entropy $n - O(n^{3/4})$. Our proofs establish a connection between the security guarantee of our protocol and the well studied problem of Kakeya over finite fields motivated by Harmonic analysis and analytic number theory (see [7, 8] for background information on Kakeya Problem). Our protocol is quite different in nature and much simpler to implement and still yield much higher min-entropy.

1.1 Previous work

There is a large body of previous work which considers the problem of random string selection, and related problems such as a leader selection and fault-tolerant computation. We refer the reader to [10] for an overview of the literature. In this paper we assume that both parties have unlimited computational power, i.e., so called *full information model.* Several different measures for the randomness guarantee of the protocol are used in the literature. The most widely used is the (μ, ϵ)-resilience and the statistical distance from the uniform distribution. Informally a protocol is (μ, ϵ)-resilient if for every set $S \subset \{0,1\}^n$ with density μ (cardinality $\mu 2^n$), the output of the protocol is in S with probability at most ϵ. In this paper we study however another very natural randomness guarantee, namely the entropy of the resulting output distribution. There is a certain relationship between the entropy and resilience, but these parameters are not interchangeable.

In [5], Goldreich et al. constructs a protocol that is $(\mu, \sqrt{\mu})$-resilient for all $\mu > 0$. This protocol runs in $O(n)$ rounds and communicates $O(n^2)$ bits. We show that their security guarantee also implies entropy $n - O(1)$. Hence, our first protocol, that runs in $4 \log^* n$

rounds, is an improvement in the number of rounds with respect to the entropy measure over that protocol.

Sanghvi and Vadhan [10] give a protocol for every constant $\delta > 0$ that is $(\mu, \sqrt{\mu + \delta})$-resilient and that has constant statistical distance from the uniform distribution. This type of resilience essentially guarantees security only for sets of constant density. Indeed, their protocol allows the cheating party to bias the output distribution so that a particular string has a constant probability of being the output. Hence, their protocol only guarantees constant *min-entropy* and entropy $(1-\epsilon)n$ for $\epsilon > 0$. Sanghvi and Vadhan also show a lower bound $\Omega(\log^* n)$ on the number of rounds of any random selection protocol that achieves constant statistical distance from the uniform distribution. We show that entropy $n - O(1)$ implies being close to uniform distribution so the lower bound translates to our protocols.

Recently, Gradwohl et al. [6], who also considered protocols with more than 2 players, constructed for each μ a $O(\log^* n)$-round protocol that is $(\mu, O(\sqrt{\mu}))$-resilient and that uses linear communication. Our results are not completely comparable with those of [6]; the protocols of [6] only achieve entropy $n - O(\log n)$ whereas the entropy $n - O(1)$ of our protocol implies only $(\mu, O(1/\log(1/\mu)))$-resilience for all $\mu > 0$. Their $(1/n^2, O(1/n))$-resilient protocol matches our non-explicit protocol from Section 4.1 in terms of entropy but our protocol can be extended to also achieve high $(n/2 - O(\log n))$ min-entropy at the cost of additional 3 rounds.

This extensibility comes from the fact that all our protocols are asymmetric. When Bob is honest (and Alice dishonest) the min-entropy of the output is guaranteed to be as high as $n - O(\log n)$, which implies, by the aforementioned result of Kitaev [2] that the min-entropy is only $O(\log n)$ when Bob is dishonest (and Alice honest). The protocols of Gradwohl et al. in general do not have this feature. Whenever their protocols achieve high $(n - O(\log n))$ entropy the min-entropy is only $O(\log n)$.

Finally our explicit protocol from Section 4.2 obtains $n - O(n^{3/4})$ entropy and thus performs worse than the explicit protocol from [6], that achieves for $\mu = 1/\log n$ entropy $n - O(\log n)$. However it still has min-entropy $n/2 - O(\log n)$ outperforming [6], that only gets min-entropy $O(\log n)$.

The paper is organized as follows. In the next section we review the notion of entropy and of other measures of randomness, and we establish some relationships among them. Section 3 contains our main protocol that achieves entropy $n - O(1)$. In Section 4 we address the problem of reducing the communication complexity of our protocols.

2 Preliminaries

Let $\mathbf{Y}$ be a random variable with a finite range S. The *entropy of* $\mathbf{Y}$ is defined by:

$$H(\mathbf{Y}) = -\sum_{s \in S} \Pr[\mathbf{Y} = s] \cdot \log \Pr[\mathbf{Y} = s].$$

If for some $s \in S$, $\Pr[\mathbf{Y} = s] = 0$ then the corresponding term in the sum is considered to be zero. All logarithms are based two.

Let $\mathbf{X}, \mathbf{Y}$ be (possibly dependent) jointly distributed random variable with ranges T, S, respectively. The *entropy of* $\mathbf{Y}$ *conditional to* $\mathbf{X}$ is defined by:

$$H(\mathbf{Y}|\mathbf{X}) = \sum_{t \in T} \Pr[\mathbf{X} = t] H(\mathbf{Y}|\mathbf{X} = t),$$

where $\mathbf{Y}|\mathbf{X} = t$ stands for the random variable whose range is S and which takes outcome $s \in S$ with probability $\Pr[\mathbf{Y} = s|\mathbf{X} = t]$.

The following are basic facts about the entropy:

$$\begin{aligned}
H(f(\mathbf{Y})) &\leq H(\mathbf{Y}) \text{ for any function } f\ , && (1)\\
H(\mathbf{Y}) &\leq \log |S|, && (2)\\
H(\mathbf{Y}|\mathbf{X}) &\leq H(\mathbf{Y}), && (3)\\
H(\langle \mathbf{X}, \mathbf{Y}\rangle) &= H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X}), && (4)\\
H(\mathbf{X}) &\leq H(\langle \mathbf{X}, \mathbf{Y}\rangle) \text{ (follows from (4))}, && (5)\\
H(\langle \mathbf{X}, \mathbf{Y}\rangle) &\leq H(\mathbf{X}) + H(\mathbf{Y}) \text{ (follows from (3) and (4))}. && (6)
\end{aligned}$$

Here $\langle \mathbf{X}, \mathbf{Y}\rangle$ stands for the random variable with range $S \times T$, which takes the outcome $\langle s, t\rangle$ with probability $\Pr[\mathbf{X} = t, \mathbf{Y} = s]$. We will abbreviate $H(\langle \mathbf{X}, \mathbf{Y}\rangle)$ as $H(\mathbf{X}, \mathbf{Y})$ in the sequel.

The following corollaries of these facts are used in the sequel

1. Let Y_i be random variables with the same range S and let $\mathbf{Y}$ be obtained by picking an index $i \in \{1, \dots, n\}$ uniformly at random and then drawing a random sample according to $\mathbf{Y}_i$. Then $H(\mathbf{Y}) \geq \frac{1}{n}\sum_{i=1}^{n} H(\mathbf{Y}_i)$. (Indeed, let $\mathbf{X}$ stand for the random variable uniformly distributed in $\{1, \dots, n\}$. Then $H(\mathbf{Y}) \geq H(\mathbf{Y}|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} H(\mathbf{Y}_i)$.)

2. Let $\ell \geq 1$ be an integer and $f : S \to T$ be a function from a set S to a set T. Let $\mathbf{Y}$ be a random variable with range S. If $\forall t \in T$, $|f^{-1}(t)| \leq \ell$ then $H(f(\mathbf{Y})) \geq H(\mathbf{Y}) - \log \ell$. (Indeed, let $\mathbf{X}$ be the index of $\mathbf{Y}$ in $f^{-1}(\mathbf{Y})$. Then $H(\mathbf{Y}) = H(f(\mathbf{Y}), \mathbf{X}) \leq H(f(\mathbf{Y})) + H(\mathbf{X}) \leq H(f(\mathbf{Y})) + \log \ell$).

The *min-entropy* of a random variable $\mathbf{X}$ with a finite range S is

$$H_\infty(\mathbf{X}) = \min\{-\log \Pr[\mathbf{X} = s] : s \in S\}.$$

It is straightforward that Shannon entropy is always greater than or equal to min-entropy:

$$H(\mathbf{X}) \geq H_\infty(\mathbf{X}).$$

The *statistical distance* between random variables $\mathbf{X}, \mathbf{Y}$ with the same finite range S is defined as the maximum

$$|\Pr[\mathbf{X} \in A] - \Pr[\mathbf{Y} \in A]|$$

over all subsets A of S. It is easy to see that the maximum is attained for A consisting of all s with $\Pr[\mathbf{X} = s] > \Pr[\mathbf{Y} = s]$ (as well as for its complement). For every integer $n \geq 1$, we denote by $\mathbf{U}_n$ the uniform probability distribution of strings $\{0, 1\}^n$.

Definition. Let r, n be natural numbers. A deterministic strategy of a player (Alice or Bob) is a function that maps each tuple $\langle x_1, \dots, x_i\rangle$ of binary strings where $i < r$ to a binary string (the current message of the player provided $\langle x_1, \dots, x_i\rangle$ is the sequence of previous messages). A randomized strategy of a player (Alice or Bob) is a probability distribution over deterministic strategies.

A *protocol running in r rounds* is a function f that maps each r-tuple $\langle x_1, \dots, x_r\rangle$ of binary strings to a binary string of length n (the first string x_1 is considered as Alice's message, the second string x_2 as Bob's message and so on) and a pair $\langle \mathbf{S}_A, \mathbf{S}_B \rangle$ of randomized strategies.

If S_A, S_B are deterministic strategies of Alice and Bob then the outcome of the protocol for S_A, S_B is defined as $f(x_1, \ldots, x_r)$ where $x_1, \ldots, x_r$ are defined recursively: $x_{2i+1} = S_A(\langle x_1, \ldots, x_{2i} \rangle)$ and $x_{2i+2} = S_B(\langle x_1, \ldots, x_{2i+1} \rangle)$.

If $\mathbf{S}_A, \mathbf{S}_B$ are randomized strategies of Alice and Bob then the outcome of the protocol is a random variable generated as follows: select independently Alice's and Bob's strategies S_A, S_B with respect to probability distributions $\mathbf{S}_A$ and $\mathbf{S}_B$, respectively, and output the result of the protocol for S_A, S_B.

We say that Alice follows the protocol (is *honest*) if she uses the strategy $\mathbf{S}_A$. We say that Alice deviates from the protocol (*cheats*) if she uses any other randomized strategy. Similarly for Bob.

We say that a protocol P for random string selection is (k,l)-*good* if the following properties hold:

1. If Alice follows the protocol and Bob deviates from it then the outcome has entropy at least k.

2. If Bob follows the protocol and Alice deviates from it then the outcome has entropy at least l.

(End of Definition.)

Throughout the paper we use the following easy observation that holds for every protocol:

Lemma 1 *Assume that Alice's strategy* $\mathbf{A}$ *guarantees that the entropy of the outcome is at least* α *for all deterministic strategies of Bob. Then the same guarantee holds for all randomized strategies of Bob as well. A similar statement is true for min-entropy in place of entropy.*

Proof. We first prove the min-entropy part. Assume that Alice's strategy $\mathbf{A}$ guarantees that the min-entropy of the outcome is at least α for all deterministic strategies B of Bob. Let $\mathbf{X}_B$ denote the outcome random variable provided Bob uses a deterministic strategy B. Then for every x the probability $\Pr[\mathbf{X}_B = x]$ is at most $2^{-\alpha}$.

Assume that Bob uses a randomized strategy $\mathbf{B}$, which is a probability distribution over his deterministic strategies, and let $\mathbf{X}$ denote the output random variable. Then $\Pr[\mathbf{X} = x]$ is equal to the average value of $\Pr[\mathbf{X}_B = x]$ with respect to that distribution. Hence the min-entropy part follows from the fact that the average value of any random variable cannot exceed its maximal value, which is at most $2^{-\alpha}$ in our case.

The Shannon entropy part follows from the inequality $H(\mathbf{X}) \geq H(\mathbf{X}|\mathbf{B})$. Indeed, $H(\mathbf{X}|\mathbf{B})$ is the average value of $H(X_B)$ over a randomly chosen B. □

For $\mu, \epsilon > 0$, a random string selection protocol P is (μ, ϵ)-*resilient* if for any set S of size at most $\mu 2^n$, the probability that the output of P is in S is at most ϵ, even if one of the parties cheats.

In order to compare our results with previous work we state the following claim.

Lemma 2 *For a random selection protocol* P *the following holds.*

1. *If* P *is* $(\mu, d\mu^c)$*-resilient for some constants* $c, d > 0$ *and any* $\mu > 0$ *then* P *is* $(n - O(1), n - O(1))$*-good.*

2. *If P is $(n - O(1), n - O(1))$-good then for some constant d and any $\mu > 0$ it is $(\mu, d/\log(1/\mu)))$-resilient.*

Proof. In order to prove the first part of the claim it suffices to show that a random variable $\mathbf{X}$ with the property that for any set S, $\Pr[\mathbf{X} \in S] < O(|S|^c/2^{cn})$ has entropy $n - O(1)$. For $x \in \{0,1\}^n$, let $p_x = \Pr[\mathbf{X} = x]$. For any integer $i < n$, define

$$S_i = \{x \in \{0,1\}^n \mid 2^{-n+i} < p_x \leq 2^{-n+i+1}\}.$$

It is straightforward that

$$H(\mathbf{X}) = -\sum_x p_x \log p_x \geq \sum_{i<n} \sum_{x \in S_i} p_x(n-i-1) = n - \sum_{i<n} \sum_{x \in S_i} p_x(i+1).$$

Since the total probability sums to one, we have $|S_i| < 2^{n-i}$ for $i < n$. As P is resilient $\sum_{x \in S_i} p_x = O(2^{-ci})$. Hence,

$$H(\mathbf{X}) \geq n - \sum_{0 \leq i < n} (i+1)2^{-ci} \geq n - O(1).$$

The second part is a corollary of the following observation: if $H(\mathbf{X}) \geq n - c$ then for every set S of density μ, we have $\Pr[\mathbf{X} \in S] \leq (c+1)/\log(1/\mu)$. Indeed, let p denote this probability. Which random variable that falls in S with probability p has maximal entropy? It is the random variable that takes every value in S with probability $p/|S|$ and takes every value outside S with probability $(1-p)/(2^n - |S|)$. The entropy of this variable equals

$$\begin{aligned}
&p \log \mu 2^n/p + (1-p)\log(1-\mu)2^n/(1-p) \\
&\leq p \log \mu 2^n + (1-p)\log(1-\mu)2^n + 1 \\
&= n + p\log\mu + (1-p)\log(1-\mu) + 1 \leq n + p\log\mu + 1.
\end{aligned}$$

Since $H(\mathbf{X}) \geq n - c$ we have $p \log \mu \geq -c - 1$ and $p \leq (c+1)/\log(1/\mu)$. □

3 The main protocol

In this section we construct a protocol that is $(n - O(1), n - O(1))$-good. We start with the following protocol.

Lemma 3 *There is a $(n-1, n-\log n)$-good protocol P_0 running in 3 rounds and communicating $n^2 + n + \log n$ bits. If Bob is honest then the outcome of P_0 has min-entropy at least $n - \log n$.*

Proof. The protocol $P_0(A, B)$ is as follows:

1. Player A picks $x_1, x_2, \ldots, x_n \in \{0,1\}^n$ uniformly at random and sends them to Player B.

2. Player B picks $y \in \{0,1\}^n$ uniformly at random and sends it to Player A.

3. Player A picks an index $j \in \{1, \ldots, n\}$ uniformly at random and sends it to B.

4. The outcome $\mathbf{R}$ of the protocol is $x_j \oplus y$, i.e., the bit-wise xor of x_j and y.

Note that the entropy bounds are tight as a cheating Bob can set $y = x_1$ in the protocol and then $H(\mathbf{R}) = n - 1$. Similarly, a cheating Alice can enforce the first $\log n$ bits of the outcome to be all zero bits so $H(\mathbf{R}) = n - \log n$ in that case.

1) Assume that Alice follows the protocol and Bob is trying to cheat. Hence, Alice picks uniformly at random $x_1, \ldots, x_n \in \{0,1\}^n$. Bob picks y. Then Alice picks a random index $j \in \{1, \ldots n\}$ and they set $\mathbf{R} = x_j \oplus y$. Clearly, $H(x_1, \ldots, x_n) = n^2$, thus

$$\begin{aligned} n^2 &= H(x_1, \ldots, x_n) \leq H(x_1, \ldots, x_n, y) \leq H(x_1 \oplus y, \ldots, x_n \oplus y) + H(y) \\ &\leq H(x_1 \oplus y, \ldots, x_n \oplus y) + n. \end{aligned}$$

Here the first inequality holds by (5), the middle one by (1) and (6), and the last one by (2). Therefore,

$$(n^2 - n)/n \leq H(x_1 \oplus y, \ldots, x_n \oplus y)/n \leq \sum_{i=1}^{n} H(x_i \oplus y)/n = H(x_j \oplus y | j) \leq H(x_j \oplus y).$$

Here the second inequality holds by (6), the equality holds, as Alice chooses j uniformly, and the last inequality is true by (3).

2) Assume that Bob follows the protocol and Alice is trying to cheat. As Shannon entropy is greater than or equal to the min-entropy, it suffices to prove the lower bound on the min-entropy. WLOG we can assume that Alice uses a deterministic strategy. Fix a deterministic strategy of Alice, which picks a particular sequence $x_1, \ldots, x_n$ in the first round and then sends a $i = i(y)$ in the third round. For every n bit string s the probability of event $x_{i(y)} \oplus y = s$ does not exceed the probability of event $\exists i,\ x_i \oplus y = s$, which is at most $n2^{-n}$ by union bound over i's. □

Our protocol achieves our goal of having entropy of the outcome close to n if Alice is honest. However if she is dishonest she can fix up-to $\log n$ bits of the outcome to her will. Clearly, Alice's cheating power comes from the fact that she can choose up-to $\log n$ bits in the last round of the protocol. If we would reduce the number of strings x_j she can choose from in the last round, her cheating ability would decrease as well. Unfortunately, that would increase cheating ability of Bob. Hence, there is a trade-off between cheating ability of Alice and Bob. To overcome this we will reduce the number of strings Alice can choose from but at the same time we will also limit Bob's cheating ability by replacing his y by an outcome of yet another run of the protocol played with Alice's and Bob's roles reversed. By iterating this several times we can obtain the following protocol.

Let $\log^* n$ stand for the number of times we can apply the function $\lceil \log x \rceil$ until we get 1 from n. For instance, $\log^* 100 = 4$.

Theorem 4 *There is a $(n-2, n-3)$-good protocol running in $2\log^* n + 1$ rounds and communicating $n^2 + O(n \log n)$ bits. Depending on n, either if Alice or Bob is honest then the min-entropy of the protocol is at least $n - O(\log n)$.*

Proof. Let $k = \log^* n - 1$. Define $\ell_0 = n$ and $\ell_i = \lceil \log \ell_{i-1} \rceil$, for $i = 1, \ldots, k$, so $\ell_{k-1} \in \{3, 4\}$ and $\ell_k = 2$.

For $i = 1, \ldots, k$ we define protocol $P_i(A, B)$ as follows.

1. Player A picks $x_1, x_2, \ldots, x_{\ell_i} \in \{0,1\}^n$ uniformly at random and sends them to Player B.
2. Players A and B now run protocol $P_{i-1}(B, A)$ (note that players exchange their roles) and set y to the outcome of that protocol.
3. Player A picks an index $j \in \{1, \ldots, \ell_i\}$ uniformly at random and sends it to B.
4. The outcome $\mathbf{R}_i$ of this protocol is $x_j \oplus y$.

We claim that the protocols are $(n-2, n-\log 4\ell_i)$-good:

Lemma 5 *For all $i = 0, 1, \ldots, k$ the following is true.*

1. *If Alice follows the protocol $P_i(\text{Alice}, \text{Bob})$ then the outcome $\mathbf{R}_i$ satisfies $H(\mathbf{R}_i) \geq n-2$.*
2. *If Bob follows the protocol $P_i(\text{Alice}, \text{Bob})$ then the outcome $\mathbf{R}_i$ of the protocol satisfies $H(\mathbf{R}_i) \geq n - \log 4\ell_i$.*

Furthermore, if i is even and Bob is honest or i is odd and Alice is honest then $H_\infty(\mathbf{R}_i) \geq n - \sum_{j=0}^{i} \log \ell_j$.

Proof. We prove both claims simultaneously by an induction on i. For $i = 0$ the claims follow from Lemma 3. So assume that the claims are true for $i-1$ and we will prove them for i.

1) If Alice follows the protocol $P_i(Alice, Bob)$ then she picks $x_1, \ldots, x_{\ell_i}$ uniformly at random. Then the protocol $P_{i-1}(Bob, Alice)$ is invoked to obtain $y = \mathbf{R}_{i-1}$. We can reason just as in the proof of Lemma 3. However this time we have a better lower bound for $H(x_1, \ldots, x_{\ell_i}, y)$. Indeed, by induction hypothesis, since Alice follows the protocol,

$$H(y|x_1, \ldots, x_{\ell_i}) \geq n - \log 4\ell_{i-1} \geq n - 2\ell_i.$$

Here the last inequality holds for all $i < k$ as $\ell_{i-1} > 4$ in this case and hence $2\ell_i \geq 2\log \ell_{i-1} > \log 4\ell_{i-1}$. For $i = k$ we have $\ell_{i-1} \in \{3, 4\}$ and $\ell_i = 2$ and the inequality is evident.

Thus,

$$H(x_1, \ldots, x_{\ell_i}, y) = H(x_1, \ldots, x_{\ell_i}) + H(y|x_1, \ldots, x_{\ell_i}) \geq \ell_i n - 2\ell_i + n.$$

Just as in Lemma 3, this implies

$$\begin{aligned} H(x_j \oplus y) \geq H(x_j \oplus y|j) &= \sum_{s=1}^{l_i} H(x_s \oplus y)/\ell_i \\ &\geq (H(x_1, \ldots, x_{\ell_i}, y) - H(y))/\ell_i \geq (\ell_i n - 2\ell_i + n - n)/\ell_i = n-2. \end{aligned}$$

2) Assume that Bob follows the protocol $P_i(Alice, Bob)$ but Alice deviates from it by carefully choosing $x_1, \ldots, x_{\ell_i}$ and j. Then the protocol $P_{i-1}(Bob, Alice)$ is invoked to obtain $y = \mathbf{R}_{i-1}$. By induction hypothesis $H(y|x_1, \ldots, x_{\ell_i}) \geq n-2$. Now Alice chooses $j \in \{1, \ldots, \ell_i\}$ and we have

$$\begin{aligned} H(x_j \oplus y) \geq H(x_j \oplus y|\langle x_1, \ldots, x_{\ell_i}\rangle) &\geq H(y|\langle x_1, \ldots, x_{\ell_i}\rangle) - H(j|\langle x_1, \ldots, x_{\ell_i}\rangle) \\ &\geq H(y|\langle x_1, \ldots, x_{\ell_i}\rangle) - H(j) \geq n - 2 - \log \ell_i. \end{aligned}$$

The claim about min-entropy follows by induction. The base of induction $i = 0$ holds by Lemma 3. The induction step: assume that i is even and Bob is honest. By induction hypothesis for the selected random string y we have $H_\infty(y) \geq n - \sum_{k=0}^{i-1} \log \ell_k$. That is, for every n bit string s,

$$\Pr[y = s] \leq 2^{-(n - \sum_{k=0}^{i-1} \log \ell_k)}.$$

By union bound for every s the probability that the outcome equals s is

$$\Pr[x_j \oplus y = s] \leq \sum_j \Pr[y = x_j \oplus s] \leq \ell_i 2^{-(n - \sum_{k=0}^{i-1} \log \ell_k)} = 2^{-(n - \sum_{k=0}^{i} \log \ell_k)}.$$

If i is odd and Alice is honest then the arguments are even simpler:

$$\Pr[x_j \oplus y = s] = (1/\ell_j) \sum_j \Pr[y = x_j \oplus s] \leq 2^{-(n - \sum_{k=0}^{i-1} \log \ell_k)} < 2^{-(n - \sum_{k=0}^{i} \log \ell_k)}.$$

□

By the lemma, the protocol P_k is $(n-2, n-3)$ good. It runs in $2k+3 = 2(\log^* n - 1) + 3$ rounds.

The number of communicated bits is equal to

$$n^2 + n + \log n + \sum_{i=1}^{k} (n\ell_i + \log \ell_i)$$

All ℓ_i's in the sum are at most $\log n$ and decrease faster than a geometric progression. Hence the sum is at most its largest term $(n \log n)$ times a constant. □

In [10], Sanghvi and Vadhan establish that any protocol for random selection that guarantees a constant statistical distance of the output from the uniform distribution requires at least $\Omega(\log^* n)$ rounds. Hence we obtain the following corollary to the following lemma, which establishes a relation between entropy and constant statistical distance.

Lemma 6 *For every real c there is a real $q < 1$ such that the following holds. If $\mathbf{X}$ is a random variable with range $\{0,1\}^n$ and $H(\mathbf{X}) \geq n - c$ then the statistical distance of $\mathbf{X}$ and $\mathbf{U}_n$ is at most q.*

We prove this lemma in Appendix.

Corollary 7 *If P is a protocol that is $(n-O(1), n-O(1))$-good then P has at least $\Omega(\log^* n)$ rounds.*

4 Improving communication complexity

In the previous section we have shown a protocol for Alice and Bob that guarantees that the entropy of the selected string is at least $n - O(1)$. The protocol has an optimal (up-to a constant factor) number of rounds and communicates $O(n^2)$ bits. In this section we will address the possibility of reducing the amount of communication in the protocol.

We focus on the basic protocol $P_0(A, B)$ as that protocol contributes to the communication the most. The protocol can be viewed as follows.

1. Player A picks $x \in \{0,1\}^{m_A}$ uniformly at random and sends it to Player B.
2. Player B picks $y \in \{0,1\}^{m_B}$ uniformly at random and sends it to Player A.
3. Player A picks an index $j \in \{0,1\}^{m'_A}$ uniformly at random and sends it to B.
4. A fixed function $f : \{0,1\}^{m_A} \times \{0,1\}^{m_B} \times \{0,1\}^{m'_A} \to \{0,1\}^n$ is applied to x, y and j to obtain the outcome $f(x,y,j)$.

We will denote such a protocol by $P_0(A,B,f)$. In the basic protocol the parameters are: $m_A = n^2$, $m_B = n$ and $m'_A = \log n$. We would like to find another suitable function f with a smaller domain.

We note first that three rounds in the protocol are necessary in order to obtain the required guarantees on the output of the protocol. By a result of [11], in any two round protocol at least one of the parties can force the output to have entropy at most $n/2 + O(\log\log n)$. (In a two round protocol, if for some x, the range of $f(x,\cdot)$ is small then Alice can enforce entropy $n/2 + \log n$ by picking this x. On the other hand, if $f(x,\cdot)$ has a large range for all x, then Bob can cheat by almost always enforcing the output to lie in a small set that intersects images of almost all functions $f(x,\cdot)$. Bob's *cheating* set can be picked at random.)

4.1 Non-explicit protocol

The following claim indicates that finding a suitable function f should be feasible.

Lemma 8 *If $f : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^{5\log n} \to \{0,1\}^n$ is taken uniformly at random among all functions then with probability at least $1/2$, $P_0(A,B,f)$ satisfies:*

1. *If Alice follows the protocol $P_0(Alice, Bob, f)$ then the outcome $\mathbf{R}$ satisfies $H(\mathbf{R}) \geq n - O(1)$.*
2. *If Bob follows the protocol $P_0(Alice, Bob, f)$ then the outcome $\mathbf{R}$ of the protocol satisfies $H(\mathbf{R}) \geq H_\infty(\mathbf{R}) \geq n - O(\log n)$.*

Proof. We will define certain properties of a function f and we will show that there are many functions with such properties and that any such function satisfies the lemma. The latter will be done in the same manner, as in the first item of Lemma 2. We have proved there that for any random variable $\mathbf{R}$ in the set $\{0,1\}^n$ we have

$$H(\mathbf{R}) \geq n - \sum_{0 \leq i < n} p_i(i+1),$$

where $p_i = \sum_{r \in S_i} p_r$, $p_r = \Pr[\mathbf{R} = r]$ and

$$S_i = \{r \in \{0,1\}^n \mid 2^{-n+i} < p_r \leq 2^{-n+i+1}\}.$$

Also we have seen that $|S_i| < 2^{n-i}$.

The properties of a function f will ensure that the outcome $f(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ of the protocol falls into any set S of size less than 2^{n-i} with probability about 2^{-i} or less. More specifically, let $K = \{0,1\}^n$ and $L = \{0,1\}^{5\log n}$. The properties of $f : K \times K \times L \to K$ are as follows:

1. For any $S \subseteq \{0,1\}^n$ of size at least $2^n/n^2$, and for any function $x \mapsto y(x)$ from K to K
$$\mathrm{Pr}_{x\in K, z\in L}[f(x, y(x), z) \in S] \le 2|S|/2^n,$$
2. For any $S \subseteq \{0,1\}^n$ of size at most $2^n/n^2$, and for any function $x \mapsto y(x)$ from K to K
$$\mathrm{Pr}_{x\in K, z\in L}[f(x, y(x), z) \in S] \le 2/n^2,$$
3. For every $s \in \{0,1\}^n$ and any $x \in K$,
$$\mathrm{Pr}_{y\in K}[s \in f(x, y, L)] \le 2n^5/2^n.$$

The first two conditions imply that the entropy of the outcome is $n - O(1)$ in the case when Alice follows the protocol (and Bob may be cheating). Indeed, w.l.o.g. we may assume that Bob uses a deterministic strategy, specified by a function $x \mapsto y(x)$. The first two conditions ensure that the sum $\sum_{0\le i<n} p_i(i+1)$ is bounded by a constant. Indeed, we can split the sum $\sum_{0\le i<n} p_i(i+1)$ into two sums: the sum over i's with $|S_i| \le 2^n/n^2$ and the rest. In the first sum $p_i \le 2/n^2$ for all i (Property 2) and hence the sum is bounded by $n \times n \times 2/n^2 = 2$. In the second sum $p_i \le 2|S_i|/2^n < 2 \cdot 2^{n-i}/2^n$ for all i (Property 1) and hence the sum is at most $\sum_{0\le i<n} 2 \cdot 2^{-i}(i+1) = O(1)$.

The bound on min-entropy follows immediately from the last property of the function f.

It remains to prove that there if a function with properties 1–3. We show that for n large enough the probability that a random function satisfies each of the properties is at least 99/100, hence a random function satisfies all of them with probability at least 97/100. To this end we will use the Chernoff bound in the following two forms. Assume that we are given independent random variables $\mathbf{T}_1, \dots, \mathbf{T}_k$ with values 0,1. Then the probability that their sum $\mathbf{T}$ exceeds twice the expectation $\mathrm{E}\mathbf{T}$ of $\mathbf{T}$ is less than $2^{-\mathrm{E}\mathbf{T}/4}$ [1, Cor A.1.14] and the probability that $\mathbf{T}$ exceeds $\mathrm{E}\mathbf{T} + \alpha k$ is less than $2^{-2\alpha^2 k}$ [1, Thm A.1.4].

1. Let $S \subseteq \{0,1\}^n$ be of size at least $2^n/n^2$ and $x \mapsto y(x)$ any mapping from K to K. The expected size of the set $\{(x,z) \mid f(x, y(x), z) \in S\}$ is $n^5|S|$. The first property of the function f claims that the size of this set exceeds the expected size at most twice. Thus by Chernoff bound the property does not hold for particular S and particular mapping $x \mapsto y(x)$ with probability at most $e^{-n^5|S|/4} \le e^{-n^3 2^n/4}$. By union bound over S and mappings $x \mapsto y(x)$ Property 1 does not hold with probability at most $2^{2^n} \cdot 2^{n2^n} \cdot e^{-n^3 2^n/4}$, which is negligible.

2. For the second property we cannot use the first form of Chernoff bound, as $|S|$ can be very small. This time we use the second form and conclude that the property does not hold for particular S and particular mapping $x \mapsto y(x)$ with probability at most $e^{-2(1/n^2)^2 n^5 2^n} = e^{-2n2^n}$. By union bound over S and mappings $x \mapsto y(x)$ Property 2 does not hold with probability at most $2^{2^n} \cdot 2^{n2^n} \cdot e^{-2n2^n}$, which is negligible.

3. For the last condition, for any s and $x \in K$, for a random function f, the expected number of occurrences of s in $f(x, K, L)$ is n^5 (it is convenient to view $f(x, K, L)$ as multiset of size $2^n n^5$). Thus by Chernoff bound, $\mathrm{Pr}_f[$ the number of occurrences of s in $f(x, K, L) > 2n^5] \le e^{-n^5/4}$. By a union bound over x and s, with a probability at least $1 - 2^{2n} \cdot e^{-n^5/4}$ for a random f, the number of occurrences of any particular element in $f(x, K, L)$ is bounded by $2n^5$, for each x. The property follows. □

The question is how to find an explicit function f of similar properties. We propose the following three functions that we believe have the required properties. We prove several results in that direction.

1. $f_{\text{rot}} : \{0,1\}^n \times \{0,1\}^n \times \{0,\dots,n-1\} \to \{0,1\}^n$ defined by $f(x,y,j) = x^j \oplus y$, where x^j is the j-th rotation of x, $x^j = x_{j+1}\cdots x_n x_1 \cdots x_j$.
2. $f_{\text{lin}} : F^{k-1} \times F^k \times F \to F^k$, where $F = GF[2^{\log n}]$, $k = n/\log n$ and $f(d,y,j) = (1,d_1,\dots,d_{k-1}) * j + (y_1,\dots,y_k)$.
3. $f_{\text{mul}} : F \times F \times H \to F$, where $F = GF[2^n]$, $H \subseteq F$, $|H| = n$, and $f(x,y,j) = x * j + y$.

In particular the function f_{rot} is interesting as it would allow very efficient implementation. We conjecture that for $f \in \{f_{\text{rot}}, f_{\text{lin}}, f_{\text{mul}}\}$ protocol $P_0(A,B,f)$ is $(n-o(n), n-o(n))$-good.

Lemma 9 *$P_0(A,B,f_{\text{rot}})$ is $(n/2-1/2, n-\log n)$-good when $n \geq 2$ is prime and the min-entropy of the outcome is at least $n - \log n$ when Bob follows the protocol.*

Proof. If Bob follows the protocol then by analysis similar to that in the proof of Lemma 3 the outcome of the protocol has min-entropy at least $n - \log n$.

So we only consider the case when Alice follows the protocol but Bob deviates from it. WLOG Bob is deterministic. Let $x \in \{0,1\}^n$ be chosen uniformly at random and $y \in \{0,1\}^n$ be set depending on x. As we have seen in the proof of Lemma 3, the entropy of the outcome is at least $\sum_{j=1}^{n} H(x^j \oplus y)/n$. Obviously, the arithmetic mean of any $n \geq 2$ numbers $a_1,\dots,a_n$ is greater than or equal to the arithmetic mean of the two least numbers among $a_1,\dots,a_n$. Thus it suffices to show that for all $i \neq j \in \{0,\dots,n-1\}$ it holds $H(x^i \oplus y) + H(x^j \oplus y) \geq n-1$.

Fix $i \neq j \in \{0,\dots,n-1\}$. We have $H(x^i \oplus y, x^j \oplus y) \geq H(x^i \oplus y \oplus x^j \oplus y) = H(x^i \oplus x^j)$. Consider x, x^i, x^j as n-dimensional vectors over $GF[2]$. Then $x^i \oplus x^j = A_{i,j}x$ for some matrix $A_{i,j} \in \{0,1\}^{n \times n}$ that depends only on i and j.

We claim that since n is prime, $A_{i,j}$ has rank $n-1$. Indeed, $A_{i,j}$ can be obtained from $A_{0,j-i}$ by permuting its columns, thus WLOG we may assume that $i = 0$. The matrix $A_{0,j}$ has two 1's in each row and each column and the remaining entries are zeros, hence the sum of all its rows is an all 0 vector. This implies that $\text{rk}(A_{0,j}) \leq n-1$. To prove the converse inequality, it suffices to show that applying to $A_{0,j}$ a linear transformation of rows we are able to obtain a matrix whose first $n-1$ rows form a triangular matrix. Notice that kth row of $A_{0,j}$ has 1s in kth and $(k+j \bmod n)$th cells (and 0s in the remaining cells). Thus the sum of kth and $(k+j \bmod n)$th rows has 1s in kth and $(k+2j \bmod n)$th cells (and 0s in the remaining cells). Similarly, the sum of kth, $(k+j \bmod n)$th and $(k+2j \bmod n)$th rows has 1s in kth and $(k+3j \bmod n)$th rows. And so on. As n is prime the equation $(k+xj) \bmod n = 0$ has a solution for all $k = 1,\dots,n$. It follows, that for every $k \neq n$ we can obtain a row that has 1's in kth and nth cells (and 0s in the remaining cells) summing certain rows of $A_{0,j}$. Thus we can obtain a matrix of rank $n-1$ applying a linear transformation to $A_{0,j}$.

As the mapping $x \mapsto A_{i,j}x$ is a homomorphism, each vector in its image has exactly 2 pre-images. Thus the mapping generates the uniform distribution over an $n-1$ dimensional space. Hence, $H(x^i \oplus x^j) = n-1$. The lemma follows. □

A similar lemma holds also for our other two candidate functions.

4.1.1 Averaging the asymmetry

One of the interesting features of our protocols is the asymmetry of cheating power of the two parties. We used this asymmetry to build the protocol with entropy $n - O(1)$. One can also use this asymmetry for "*averaging*" their cheating powers in the following simple way. Given a protocol $Q_n(A, B)$ for selecting an n bit string, Alice and Bob first select the first $n/2$ bits of the string by running the protocol $Q_{n/2}(\text{Alice}, \text{Bob})$ and then they select the other half of the string by running the protocol $Q_{n/2}(\text{Bob}, \text{Alice})$.

Lemma 10 *If the protocol Q_n is $(k(n), l(n))$-good then the* averaging *protocol is $(k(n/2) + l(n/2), k(n/2) + l(n/2))$-good. Similarly if the min-entropy when Alice follows the protocol is bounded from below by $k_\infty(n)$ and when Bob follows the protocol by $l_\infty(n)$, then the min-entropy of the outcome of the averaging protocol is at least $k_\infty(n/2) + l_\infty(n/2)$.*

Proof. Assume that Alice is honest and hence follows the strategy A prescribed by the protocol $Q_{n/2}(\text{Alice}, \text{Bob})$ to select the first half of the output and the strategy B prescribed by the protocol $Q_{n/2}(\text{Bob}, \text{Alice})$ to select the second half of the output. To prove the first statement, we have to show that whatever strategy S follows Bob, the entropy of the outcome $\mathbf{X}$ is at least $k(n/2) + l(n/2)$. By Lemma 1 we may assume that S is deterministic.

Let $\mathbf{X}_1, \mathbf{X}_2$ denote the first and the second part of the output, respectively. Then

$$H(\mathbf{X}) = H(\mathbf{X}_1) + H(\mathbf{X}_2|\mathbf{X}_1).$$

As the protocol $Q_{n/2}(\text{Alice}, \text{Bob})$ is $(k(n/2), l(n/2))$-good we have $H(\mathbf{X}_1) \geq k(n/2)$ and it remains to show that $H(\mathbf{X}_2|\mathbf{X}_1) \geq l(n/2)$. By definition, $H(\mathbf{X}_2|\mathbf{X}_1)$ is the average of $H(\mathbf{X}_2|\mathbf{X}_1 = x_1)$ over all outcomes x_1 of $\mathbf{X}_1$. As $\mathbf{X}_1$ is a function of messages $\mathbf{M}_1$ sent while selecting $\mathbf{X}_1$, the conditional entropy $H(\mathbf{X}_2|\mathbf{X}_1)$ also equals the average of $H(\mathbf{X}_2|\mathbf{M}_1 = m_1)$ over all possible messages m_1 sent while selecting $\mathbf{X}_1$. As the protocol $Q_{n/2}(\text{Bob}, \text{Alice})$ is $(l(n/2), k(n/2))$-good, for every m_1 we have $H(\mathbf{X}_2|\mathbf{M}_1 = m_1) \geq l(n/2)$. Indeed, once we fix m_1, the action of Bob's strategy S while selecting the second half of the output becomes deterministic.

The bound on min-entropy is proven in a similar way: for all x_1, x_2 we have

$$\Pr[\mathbf{X} = (x_1, x_2)] = \Pr[\mathbf{X}_1 = x_1] \cdot \Pr[\mathbf{X}_2 = x_2|\mathbf{X}_1 = x_1].$$

The first factor here is at most $2^{-k_\infty(n/2)}$, as $Q_{n/2}(\text{Alice}, \text{Bob})$ guarantees min-entropy at least $k_\infty(n/2)$ provided Alice is honest. The second factor is at most $2^{-l_\infty(n/2)}$, as for all messages m_1 we have $\Pr[\mathbf{X}_2 = x_2|\mathbf{M}_1 = m_1] \geq 2^{-l_\infty(n/2)}$. □

Hence from Lemma 9 we obtain the following corollary.

Corollary 11 *There is a 6-round protocol for random string selection that communicates $2n + O(\log n)$ bits, that is $(3n/4 - O(\log n), 3n/4 - O(\log n))$-good and that has min-entropy at least $n/2 - O(\log n)$ when at least one of the parties follows the protocol.*

In the next section we present a protocol that beats this one both with respect to the number of rounds and entropy.

4.2 Geometric protocols and the problem of Kakeya

We exhibit here a variant of the protocol $P_0(A, B, f_{\text{lin}})$ and show that it achieves entropy at least $n - o(n)$ if at least one party is honest. Fix a finite field F and a natural $m \geq 2$. Let $q = |F|$. We rephrase the protocol as follows:

1. Alice picks at random a vector $d = (1, d_2, \dots, d_m) \in F^m$ and sends it to Bob.
2. Bob picks at random $x = (x_1, \dots, x_m) \in F^m$ and sends it to Alice.
3. Alice picks at random $t \in F$ and sends it to Bob.
4. The output of the protocol is

$$y = x + td = (x_1 + t, x_2 + td_2, \dots, x_m + td_m).$$

The geometric meaning of the protocol is as follows. Alice picks at random a direction of an affine line in the m-dimensional space F^m over F. Bob chooses a random affine line going in that direction. Alice outputs a random point lying on the line.

It is easy to lower bound the entropy of the output y of this protocol assuming that Bob is honest.

Lemma 12 *If Bob is honest then the outcome y of the protocol satisfies*

$$H(y) \geq H_\infty(y) \geq (m-1)\log q.$$

Note that Alice can cheat this much. For example, Alice can force $y_1 = 0$ by choosing always $t = -x_1$.

Proof. Fix an outcome $a \in F^m$. Then by union bound

$$\Pr[x + td = a] \leq \sum_{s \in F} \Pr[x + sd = a] = q \cdot q^{-m}.$$

□

In the case when Alice is honest we are able to prove the bound $H(y) \geq m \log q - O(m^3)$. This question is related to the following problem known as Kakeya problem for finite fields.

Kakeya problem. *Let L be a collection of affine lines in F^m such that for each direction there is exactly one line in L going in that direction. Let P_L denote points in lines from L. How small can be $|P_L|$?*

Call any set of lines L satisfying the conditions of Kakeya problem a Kakeya family. For every Kakeya family L consider the following deterministic Bob's strategy: choose any point in the line in L going in direction d specified by Alice. Using this strategy Bob can force the outcome to be in P_L so that its entropy is at most $\log |P_L|$. Thus to prove that the entropy of the outcome is at least α (provided Alice is honest) we need the lower bound $|P_L| \geq 2^\alpha$ for Kakeya problem. Recently Dvir [3] has shown that $|P_L| \geq \binom{q+m-1}{m}$. Using the technique of [3, 4] we will show that the the entropy of the outcome of the protocol (provided Alice is honest) is at least $m \log q - O(m^3)$.

Theorem 13 *If Alice is honest then the outcome $\mathbf{Y}$ of the geometric protocol is $\varepsilon, 2m\varepsilon^{1/m}$-resilient (for all ε) and $H(Y) \geq m \log q - O(m^3)$. Moreover, $H_\infty(\mathbf{Y}) \geq \log q$.*

Proof. We start by proving that $\mathbf{Y}$ is $\varepsilon, 2m\varepsilon^{1/m}$-resilient for all ε. Assume first that Bob uses a deterministic strategy: for every d he chooses a point $x(d) \in F^m$. We have to show that the random variable $\mathbf{Y} = x(d) + dt$ is ε, δ-resilient for $\delta = 2m\varepsilon^{1/m}$. For the sake of contradiction assume that there is $T \subset F^m$ with $T = \varepsilon q^m$ and $\Pr[x(d) + dt \in T] > \delta$. Let us find a non-zero low-degree polynomial $P(x_1, \dots, x_m)$ that vanishes on T. Such polynomial can be found by solving a system of $|T|$ linear homogeneous equations. Indeed, for every $x \in T$ the condition $P(x) = 0$ is a homogeneous linear equation in the coefficients of P. We need to choose the degree k of P so that the number of coefficients of P be greater than the number of equations. Assuming that the degree of P in each variable is at most k, we are fine if $(k+1)^m > |T| = \varepsilon q^m$. Thus we can let $k = \varepsilon^{1/m} q$.

Recall that we assume that with probability more than δ it happens that $x(d) + dt \in T$ (and hence $P(x(d) + dt) = 0$). Call a direction d *good* if $x(d) + dt \in T$ with probability more than $\delta/2$ (for a random t). It is easy to see that more than $\delta/2$ fraction of directions are good.

We claim that, for all good d, the univariate polynomial $P(x(d) + dt)$ of t is identically zero. Indeed, its degree is at most km and it vanishes in more than $q\delta/2$ points. The degree of $P(x(d) + dt)$ is at most km and recall that $k = \varepsilon^{1/m} q$ and we have chosen δ so that $km = q\delta/2$.

Let $P(x) = P_0 + P_1(x) + \cdots + P_i(x)$ where P_j is a homogeneous polynomial of degree j and P_i is non-zero. The highest coefficient of the polynomial $P(x(d) + dt)$ is equal to $P_i(d)$. Thus $P_i(d) = 0$ for all good d. The Schwartz-Zippel lemma states that a n-variable non-zero polynomial over F of degree j cannot have more than $j|F|^{n-1}$ zeros. And $P_i(d)$ has $m-1$ variables and vanishes in more than $\delta q^{m-1}/2 = kmq^{m-2} \geq iq^{m-2}$ points, a contradiction. Thus we have shown that the outcome $\mathbf{Y}$ is $\varepsilon, m\varepsilon^{1/m}$-resilient provided Bob uses a deterministic strategy.

Assume now that Bob uses a randomized strategy. His strategy can be regarded as a weighted sum of deterministic strategies. Thus the outcome of the protocol is a weighted sum of $\varepsilon, m\varepsilon^{1/m}$-resilient random variables, which is $\varepsilon, m\varepsilon^{1/m}$-resilient as well.

Let us show now that the outcome $\mathbf{Y}$ of the protocol has large Shannon entropy (provided Alice is honest). For every $y \in F^m$ let $p(y) = \Pr[\mathbf{Y} = y]$ denote the probability that the result of the protocol is y. We classify all y's according to how large is y. Let S_i be the set of all y with $q^{-m}2^i \leq p(y) < q^{-m}2^{i+1}$. As in the proof of Lemma 2, one can show that

$$H(\mathbf{Y}) > m \log q - \sum_{i \geq 0} (i+1) \Pr[Y \in S_i]$$

and that $|S_i| \leq 2^{-i} q^m$. As $\mathbf{Y}$ is $2^{-i}, m2^{-i/m+1}$-resilient we obtain $\Pr[Y \in S_i] \leq m2^{-i/m+1}$. Therefore

$$\sum_{i \geq 0} (i+1) \Pr[Y \in S_i] \leq \sum_{i \geq 0} (i+1) m 2^{-i/m+1} = 2m \sum_{i \geq 0} (i+1) 2^{-i/m}.$$

We claim that the last sum is $O(m^2)$. Indeed, we can rewrite it as

$$\sum_{j=0} \sum_{i \geq j} 2^{-i/m}.$$

The inner sum $\sum_{i\ge j} 2^{-i/m}$ is the sum of geometric series with the quotient $2^{-1/m}$ and the first term $2^{-j/m}$ and thus equals

$$\frac{2^{-j/m}}{1-2^{-1/m}} = \frac{2^{-j/m}}{1-(1-\Omega(1/m))} = O(m2^{-j/m}).$$

Thus the outer sum is

$$\sum_{j=0} O(m2^{-j/m}) = O(m\sum_{j=0} 2^{-j/m}) = O(m \cdot O(m)) = O(m^2).$$

Hence $H(\mathbf{Y}) \ge m \log q - O(m^3)$. □

If we choose $m = n^{1/4}$ and $\log q = n^{3/4}$ then the lower bounds for $H(y)$ in the cases when Alice cheats and Bob cheats coincide and are equal to $n - O(n^{3/4})$. Thus we get an explicit 3 round protocol with linear communication and entropy $n - o(1)$:

Theorem 14 *There is an explicit $(n - O(n^{3/4}), n - O(n^{3/4}))$-good 3-round protocol that communicates $2n$ bits.*

Using averaging we obtain the following corollary:

Theorem 15 *There is a $(n-O(n^{3/4}), n-O(n^{3/4}))$-good 6-round protocol that communicates $2n$ bits and guarantees the min-entropy at least $n/2 - O(1)$ for both players.*

Acknowledgments

We would like to thank to Troy Lee and John Tromp for useful discussions and Navin Goyal for pointing us to the problem of Kakeya. Part of the work was done while the second, third, fourth, and sixth author were visiting CWI, Amsterdam. H. Buhrman was supported by EU project QAP and BRICKS project AFM1. H. Buhrman and M. Koucký were supported in part by an NWO VICI grant (639.023.302). M. Koucký was supported in part by grant GA ČR 201/07/P276, 201/05/0124, project No. 1M0021620808 of MŠMT ČR and Institutional Research Plan No. AV0Z10190503. The work of N. Vereshchagin was partially supported by RFBR, research project No. 12-01-00864-a.

References

[1] Noga Alon, Joel Spencer, *The probabilistic method.* John Wiley & sons, 2nd edition, 2000.

[2] Andris Ambainis, Harry Buhrman, Yevgeniy Dodis, and Hein Röhrig, Multiparty Quantum Coin Flipping. IEEE Conference on Computational Complexity 2004, pages 250–259, 2004.

[3] Z. Dvir. On the size of Kakeya sets in finite fields. J. Amer. Math. Soc., 22:1093-1097, 2009.

[4] Z. Dvir and A. Wigderson. Kakeya sets, new mergers and old extractors. In FOCS '08: Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, pages 625-633. IEEE Computer Society, 2008.

[5] O. Goldreich, S. Goldwasser, and N. Linial, Fault-tolerant computation in the full information model. SIAM Journ. on Computing 27 (2), pages 506–544, 1998.

[6] R. Gradwohl, S. Vadhan, D. Zuckerman, Random selection with an Adversarial Majority In C. Dwork, editor, Advances in Cryptology—CRYPTO '06, number 4117 in Lecture Notes in Computer Science, pages 409–426, Springer-Verlag, 20–24 August 2006. Electronic Colloquium on Computational Complexity, Technical Report TR06-026, February 2006.

[7] Gerd Mockenhaupt, Terence Tao, Restriction and Kakeya phenomena for finite fields. Duke Math. J. 121 (2004), 35–74.

[8] Thomas Wolff, Recent work connected with the Kakeya problem, in Prospects In Mathematics, H. Rossi, ed., AMS 1999.

[9] An. Muchnik and N. Vereshchagin. "Shannon Entropy vs. Kolmogorov Complexity". Computer Science — Theory and Applications: First International Computer Science Symposium in Russia, CSR 2006, St. Petersburg, Russia, June 8-12. 2006. Proceedings. Editors: Dima Grigoriev, John Harrison, Edward A. Hirsch Lecture Notes in Computer Science, vol. 3967 / 2006, pages 281–291.

[10] S. Sanghvi, S. Vadhan. The Round Complexity of Two-Party Random Selection. Thirty-seventh Annual ACM Symposium on Theory of Computing. Baltimore, MD, USA. Proceedings Pages: 338–347.

[11] T. Stepanov. Random Selection in Few Rounds. Proceedings of 8th International Computer Science Symposium in Russia, CSR 2013, Ekaterinburg, Russia, June 25-29, 2013. Lecture Notes in Computer Science v. 7913, pages 354–365.

A Appendix: The proof of Lemma 6

Fix c. For $x \in \{0,1\}^*$ let $p_x = \Pr[\mathbf{X} = x]$. The statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ is equal to $\sum_{x:p_x>2^{-n}}(p_x - 2^{-n})$. For all integer $i \le n$ let N_i stand for the number of x with

$$2^{-n+i-1} < p_x \le 2^{-n+i} \tag{7}$$

and w_i for their cumulative probability. In terms of w_i, N_i the statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ can be rewritten as

$$\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} N_i 2^{-n} \le \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} 2^{-i} w_i.$$

Here the last inequality holds, as $N_i \ge w_i 2^{n-i}$.

Thus it suffices to prove that for some $q < 1$ depending only on c it holds

$$\sum_{i=1}^{n} (1 - 2^{-i}) w_i \le q$$

provided $H(\mathbf{X}) \geq n - c$. As in the proof of Lemma 2 we can see that

$$H(\mathbf{X}) \leq \sum_{i \leq n} w_i(n+1-i) = n+1-\sum_{i \leq n} iw_i$$

hence

$$\sum_{i \leq n} iw_i \leq c+1 \tag{8}$$

Here i ranges over all integers $i \leq n$, including negative ones. However, the contribution of negative i's is bounded by a constant. Indeed, as $2^{n-i}w_i \leq N_i \leq 2^n$ we can conclude that $w_i \leq 2^i$ hence

$$0 \geq \sum_{i<0} iw_i \geq \sum_{i<0} i2^i = O(1).$$

Thus, inequality (8) implies that the sum of iw_i over positive i's is bounded by a constant:

$$\sum_{i=1}^{n} iw_i \leq c + O(1) \leftrightharpoons d.$$

Divide the sum $\sum_{i=1}^{n}(1-2^{-i})w_i$ into two groups: the sum over all $i \geq 2d$ and the rest. The first sum is small by the last inequality, as it implies $\sum_{i=2d}^{n} w_i \leq 1/2$. In the second sum the coefficient $(1-2^{-i})$ is small:

$$1-2^{-i} \leq 1-2^{-2d} \leftrightharpoons q'.$$

Thus the total sum can be upper bounded by $1/2+q'/2$, which is less than 1. Lemma 6 is proved.

Remark. It is easy to see that the above proof gives an upper bound $q = 1-2^{-2c+O(1)}$. Strengthening the arguments we can prove the lemma with $q = 1-2^{-c}+e^{-2^c}$ (see the next Lemma 16), which is nearly tight for large c. Indeed, the distance between the random variable $\mathbf{X}$ uniformly distributed over a 2^{n-c}-element set and $\mathbf{U}_n$ is $1-2^{-c}$, while $H(\mathbf{X}) = n-c$.

A converse of the proposition appears in [9], where it is proven that if the statistical difference between $\mathbf{X}$ and $\mathbf{U}_n$ is at most ε then $H(\mathbf{X}) \geq n(1-\varepsilon)-1$. That bound is also almost tight. Indeed, let $\Pr[\mathbf{X} = 0^n] = \varepsilon$ and all the other outcomes of $\mathbf{X}$ are equiprobable. The statistical distance difference between $\mathbf{X}$ and U_n is less than ε, while $H(\mathbf{X}) < n(1-\varepsilon)+1$. Hence, the constant statistical distance of $\mathbf{X}$ from $\mathbf{U}_n$ does not imply entropy $n - O(1)$.

Lemma 16 *Let $n \geq 1$ be an integer, let c be a real and $\mathbf{X}$ a random variable with range $\{0,1\}^n$. If $H(\mathbf{X}) \geq n-c$ then the statistical distance of $\mathbf{X}$ and $\mathbf{U}_n$ is at most $1-2^{-c}+e^{-2^c}$.*

Proof. For $x \in \{0,1\}^*$ let $p_x = \Pr[\mathbf{X} = x]$. For all real $z \leq n$ let N_z stand for the number of x with $p_x = 2^{n-z}$ and w_z for their total probability. We consider in the sequel only z with $w_z > 0$. The number of such z is finite. Obviously,

$$\sum_{z} w_z \leq 1. \tag{9}$$

The number N_z of x with $p_x = 2^{n-z}$ is equal to $w_z 2^{n-z}$. Thus we have

$$\sum_z w_z 2^{n-z} \le 2^n \Rightarrow \sum_z 2^{-z} w_z \le 1. \tag{10}$$

In terms of w_z the entropy of $\mathbf{X}$ is expressed as

$$H(\mathbf{X}) = \sum_z w_z(n-z) = n - \sum_z z w_z.$$

Recalling that the entropy of $\mathbf{Y}$ is at least $n-c$ we obtain

$$\sum_z z w_z \le c. \tag{11}$$

Thus is suffices to show that the statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ is at most $1 - 2^{-d} + e^{-2^d}$ for every random variable $\mathbf{X}$ satisfying (9), (10) and (11). The statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ is equal to

$$\sum_{x:p_x>2^{-n}} (p_x - 2^{-n}) = \sum_{z>0} w_z - \sum_{z>0} N_z 2^{-n} = \sum_{z>0} w_z - \sum_{z>0} 2^{-z} w_z.$$

Thus it suffices to prove that inequalities (10) and (11) imply

$$\sum_{z>0} (1 - 2^{-z}) w_z \le 1 - 2^{-d} + e^{-2^d}.$$

Note that if $w_z = 0$ for all $z \ne d$ and $w_d = 1$, then (9), (10) and (11) are true and the last sum evaluates to $1 - 2^{-d}$. Thus we need to prove that this is nearly optimal solution to the linear program

$$\sum_{z>0} w_z(1 - 2^{-z}) \to \max \qquad \text{subject to (9), (10) and (8).}$$

We apply a dual argument. Multiply inequalities (9), (10) and (11) by certain non-negative reals α, β and γ, respectively, and sum up the resulting inequalities.

The coefficients α, β and γ will be chosen so that the right hand side of the resulting inequality is equal to $1 - 2^{-d} + e^{-2^d}$ and its lefthand side is larger than $\sum_{z>0} w_z(1 - 2^{-z})$.

For every z the term w_z appears in the resulting inequality with the coefficient

$$\alpha + \beta 2^{-z} + \gamma z.$$

We have to choose α, β, γ so that for all $z \le 0$ this coefficient is non-negative and for all positive z it is at least $1 - 2^{-z}$.

Taking the derivative we can see that the minimal value of the function

$$\alpha + \beta 2^{-z} + \gamma z$$

is equal to

$$\alpha + \frac{\gamma}{\ln 2} + \gamma \log \frac{\beta \ln 2}{\gamma},$$

attained for $z = \log \frac{\beta \ln 2}{\gamma}$. Thus it suffices to have

$$0 \le \alpha + \frac{\gamma}{\ln 2} + \gamma \log \frac{\beta \ln 2}{\gamma} \tag{12}$$

and

$$1 \le \alpha + \frac{\gamma}{\ln 2} + \gamma \log \frac{(1+\beta) \ln 2}{\gamma}. \tag{13}$$

For given β, γ the best α is the minimal one satisfying (12) and (13). It is worth to choose β and γ so that the minimal α satisfying (12) coincides with the minimal α satisfying (13). This means that $\gamma \log \beta = -1 + \gamma \log(1+\beta)$, and $\beta = \frac{1}{2^{1/\gamma}-1}$.

However, for such choice of β the resulting expression for the goal function is too hard to analyze. Therefore, we decrease β a little bit and set $\beta = 2^{-1/\gamma}$. Such choice makes the right hand side of both inequalities simpler but smaller. The the right hand side of the first inequality becomes less than that of the second one. The minimal α satisfying (12) and (13) is thus equal to

$$\alpha = 1 - \frac{\gamma}{\ln 2} - \gamma \log \frac{\ln 2}{\gamma}.$$

Now it remains to choose γ minimizing the sum

$$\alpha + \beta + d\gamma = 1 - \frac{\gamma}{\ln 2} + \gamma \log \frac{\gamma}{\ln 2} + 2^{-1/\gamma} + d\gamma.$$

To simplify matters let us choose γ minimizing the sum of all the terms except $2^{-1/\gamma}$. That sum is minimal for $\gamma = 2^{-d} \ln 2$. Plugging γ into expressions for α and β we obtain

$$\alpha = 1 - 2^{-d} - d2^{-d} \ln 2, \qquad \beta = e^{-2^d}$$

and

$$\alpha + \beta + d\gamma = 1 - 2^{-d} + e^{-2^d}.$$

It remains to verify that $\alpha \ge 0$, which is straightforward. □