High Entropy Random Selection Protocols
[Extended Abstract]

Harry Buhrman[1], Matthias Christandl[2], Michal Koucký[3], Zvi Lotker[4], Boaz Patt-Shamir[5], and Nikolai Vereshchagin[6]

[1] CWI, Amsterdam, buhrman@cwi.nl
[2] Cambridge University, mc380@cam.ac.uk
[3] Institute of Mathematics of Czech Academy of Sciences, mkoucky@math.cas.cz
[4] Ben Gurion University, zvilo@cse.bgu.ac.il
[5] Tel Aviv University, boaz@eng.tau.ac.il
[6] Lomonosov Moscow State University, ver@mech.math.msu.su

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 Goldwasser et al. [3] 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. [4], however achieves much higher min-entropy: $n/2 - O(\log n)$ versus $O(\log n)$.

Finally we exhibit very simple explicit protocols. We connect the security parameter of these geometric protocols with the well studied Kakeya problem motivated by harmonic analysis and analytical number theory. We are only able to prove that these protocols have entropy $3n/4$ but still $n/2 - O(\log n)$ min-entropy. Therefore they do not perform as well with respect to the explicit constructions of Gradwohl et al. [4] entropy-wise, but still have much better min-entropy. We conjecture that these simple protocols achieve $n - o(n)$ entropy. Our geometric construction and its relation to the Kakeya problem follows a new and different approach to the random selection problem than any of the previously known protocols.

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 which follows from a result of Sanghvi and Vadhan [8]. We further consider the question of reducing the communication complexity of our protocols. 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 Goldwasser et al. [3] and from a bound on quantum coin-flipping due to Kitaev (see [2]). We propose several explicit and very simple protocols that have entropy $3n/4$ and conjecture to have entropy $n - o(n)$. Our proofs establish a connection between the security guarantee of our protocols and the well studied problem of Kakeya over finite fields motivated by Harmonic analysis and analytic number theory. Although these constructive protocols do not achieve the same parameters as the best known constructive protocols (see next section), our (geometric) protocols are 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 [8] 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 natural randomness guarantee, namely the entropy of the resulting output distribution. There is a relationship between the entropy and resilience, but these parameters are not interchangeable.

In [3], 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 uses $4 \log^* n$ is an improvement in the number of rounds with respect to the entropy measure over that protocol.

Sanghvi and Vadhan [8] present a protocol for every constant $\delta > 0$ that is $(\mu, \sqrt{\mu + \delta})$-resilient and that has constant statistical distance from the uniform distribution. They 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. [4], 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 [4]; the protocols of [4] only achieve entropy $n - O(\log n)$ and entropy $n - O(1)$ implies only $(\mu, O(1/\log(1/\mu)))$-resilience for all $\mu > 0$. Their protocol, non-explicit for $\mu = 1/n^2$ matches our non-explicit protocol from Section 4.1 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 feature 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 geometric protocol only obtains $3n/4$ entropy and thus performs worse than the explicit protocol from [4], that achieves for $\mu = 1/\log n$ entropy $n - o(n)$. Our explicit protocol though still have min-entropy $n/2 - O(\log n)$ outperforming [4], 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 protocol that achieves entropy $n - O(1)$. In Section 4 we address the problem of reducing 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:

$$H(f(\mathbf{Y})) \leq H(\mathbf{Y}) \text{ for any function } f \ , \tag{1}$$

$$H(\mathbf{Y}) \leq \log |S|, \tag{2}$$

$$H(\mathbf{Y}|\mathbf{X}) \leq H(\mathbf{Y}), \tag{3}$$

$$H(\langle \mathbf{X}, \mathbf{Y} \rangle) = H(\mathbf{Y}|\mathbf{X}) + H(\mathbf{X}), \tag{4}$$

$$H(\mathbf{X}) \leq H(\langle \mathbf{Y}, \mathbf{X} \rangle) \text{ (follows from (4))}, \tag{5}$$

$$H(\langle \mathbf{Y}, \mathbf{X} \rangle) \leq H(\mathbf{Y}) + H(\mathbf{X}) \text{ (follows from (3) and (4))}. \tag{6}$$

Here $\langle \mathbf{Y}, \mathbf{X} \rangle$ 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 $H(\langle \mathbf{Y}, \mathbf{X} \rangle)$ as $H(\mathbf{Y}, \mathbf{X})$ 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 $i \in \{1, \ldots, n\}$ 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 $\{1, \ldots, n\}$. 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

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

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 $\{0, 1\}^n$.

Let $0 < \alpha, \beta \leq 1$. We say that a random variable $\mathbf{X}$ with a range S is (α, β)-*resilient* if for every subset $S' \subseteq S$ of size $|S'| = \alpha|S|$, $\Pr[\mathbf{X} \in S] \leq \beta$.

The above four notions of randomness, i.e., entropy, min-entropy, statistical distance from uniform distribution and resilience, are not entirely comparable. In order to facilitate comparison with previous work and to apply a lower-bound from [8] to show that our main protocol needs $\Omega(\log^* n)$ rounds we establish the following relationships.

Lemma 1. *Let $\mathbf{X}$ be a random variable with a range $\{0,1\}^n$. For any constant $0 < c, c''$ and $0 < c' < 1$ there are constants d, d', d'' such that the following hold:*

1. *If for any μ, $\mathbf{X}$ is $(\mu, c\mu^{c'})$-resilient then $H(\mathbf{X}) \geq n - d$ and $H_\infty(\mathbf{X}) \geq cn - d$.*
2. *If $H(\mathbf{X}) \geq n - c''$ then for all μ, $\mathbf{X}$ is $(\mu, d'/\log(1/\mu))$-resilient and the statistical distance of $\mathbf{X}$ from U_n is at most $d'' < 1$.*

Definition. Let r, n be natural numbers. A deterministic strategy of a player (Alice or Bob) is a function that maps each tuple $\langle x_1, \ldots, x_i \rangle$ of binary strings where $i < r$ to a binary string (the current message of the player provided $\langle x_1, \ldots, x_i \rangle$ 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 $\langle x_1, \ldots, x_r \rangle$ 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 (Bob) follows the protocol if Alice (Bob) uses the strategy $\mathbf{S}_A$ ($\mathbf{S}_B$). We say that Alice (Bob) deviates from the protocol if Alice (Bob) uses any other randomized strategy.

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

1. If both Alice and Bob follow the protocol then the outcome is a fully random string of length n.
2. If Alice follows the protocol and Bob deviates from it then the outcome has entropy at least k.
3. 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 (see appendix for a proof):

Lemma 2. *Assume that Alice's strategy $\mathbf{S}_A$ 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.*

In [8], 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 previous lemma.

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

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 (see appendix for proof).

Lemma 3. *For a random selection protocol P the following holds.*

1. *If P is $(\mu, O(\mu^c))$-resilient for some constant c 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 any $\mu > 0$ it is $(\mu, 1/\log(1/\mu))$-resilient.*

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 4. *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 last two items 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) It is easy to verify that the outcome $\mathbf{R}$ of the protocol $P_0(\mathrm{Alice}, \mathrm{Bob})$ is uniformly distributed if both Alice and Bob follow the protocol and hence it has entropy n.

2) 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

$$n^2 = H(x_1, \ldots, x_n) \le H(x_1, \ldots, x_n, y) \le H(x_1 \oplus y, \ldots, x_n \oplus y) + H(y) \le H(x_1 \oplus y, \ldots, x_n \oplus y) + n.$$

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

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

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

3) Assume that Bob follows the protocol and Alice is trying to cheat. Hence, Alice carefully selects $x_1, \ldots, x_n$, Bob picks a random string $y \in \{0,1\}^n$ and Alice carefully chooses $j \in \{1, \ldots, n\}$. Thus $H(y|\langle x_1, \ldots, x_n\rangle) = n$ and hence

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

Here the second inequality holds by (1) and (6). Verifying the lower bound on the min-entropy is straightforward.

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 1. *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 both Alice and Bob follow the protocol $P_i(Alice, Bob)$ then its outcome $\mathbf{R}_i$ satisfies $H(\mathbf{R}_i) = n$.*

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

All the bounds on the entropy are valid also when conditioned on the tuple consisting of all strings communicated before running P_i. *Furthermore, if* i *is even and Bob is honest or* i *is odd and Alice is honest then* $H_\infty(\mathbf{R}_i) \geq n - \sum_{j=1}^{i+1} \ell_j$.

Proof. The first claim is straightforward to verify. We prove the other two simultaneously by an induction on i. For $i = 0$ the claims follow from Lemma 4. So assume that the claims are true for $i - 1$ and we will prove them for i.

If Alice follows the protocol $P_i(Alice, Bob)$ then she picks $x_1, \dots, x_{\ell_i}$ 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 4. However this time we have a better lower bound for $H(x_1, \dots, x_{\ell_i}, y)$. Indeed, by induction hypothesis, since Alice follows the protocol,

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

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,

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

Just as in Lemma 4, this implies

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

Assume that Bob follows the protocol $P_i(Alice, Bob)$ but Alice deviates from it (she is trying to cheat) carefully choosing $x_1, \dots, x_{\ell_i}$ and j. Then the protocol $P_{i-1}(Bob, Alice)$ is invoked to obtain $y = \mathbf{R}_{i-1}$. By induction hypothesis $H(y|x_1, \dots, x_{\ell_i}) \geq n - 2$. Now Alice chooses $j \in \{1, \dots, \ell_i\}$. Similarly as in the proof of Lemma 4, we have

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

The claim about min-entropy follows by induction.

By the lemma to 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 $\log n$ end decrease faster than a geometric progression. Hence the sum is at most its largest term ($n \log n$) times a constant.

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 smaller domain.

We note first that three rounds in the protocol are necessary in order to obtain required guarantees on the output of the protocol. In any two round protocol at least one of the parties can force the output to have entropy at most $n/2 + \log n$. (In two round protocol, if for some x, the range of $f(x, \cdot)$ is smaller than $n2^{n/2}$ 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 set of size $2^{n/2}$. 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 6. *If $f : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^{8\log n} \to \{0,1\}^n$ is taken uniformly at random among all functions then with probability at least $1/2$, $P_0(A, B, f)$ satisfies:*

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

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. $f_{\mathrm{rot}} : \{0,1\}^n \times \{0,1\}^n \times \{1, \ldots, n\} \to \{0,1\}^n$ defined by $f(x, y, j) = x^j \oplus y$, where x^j is the j-th rotation of x, $x^j = x_j x_{j+1} \cdots x_n x_1 \cdots x_{j-1}$. Here n is assumed to be a prime.

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 $f(d, y, j) = (1, d_1, \ldots, d_{k-1}) * j + (y_1, \ldots, y_k)$.
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 $(n - o(n), n - O(\log n))$-good.

Lemma 7. *$P_0(A, B, f_{\text{rot}})$ is $(n/2 - 3/2, n - \log n)$-good when n is prime and the min-entropy of the outcome is at least $n - O(\log n)$ when Bob follows the protocol.*

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

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 $Q_{n/2}(\text{Bob}, \text{Alice})$. 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 $k_\infty(n/2) + l_\infty(n/2)$.

Hence from Lemma 7 we obtain the following corollary.

Corollary 2. *There is a 5-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 show for a variant of $P_0(A, B, f_{\text{lin}})$ a similar security guarantee.

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 $3n/4 - O(1)$ 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 $d = (1, d_2, \ldots, d_m) \in F^m$ and sends it to Bob.
2. Bob picks at random $x = (x_1, \ldots, x_m) \in F^m$ 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

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

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 (see appendix for proof).

Lemma 8. *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 $t = -x_1$.

In the case when Alice is honest we are able to prove the bound $H(y) \geq (m/2 + 1)\log q - O(1)$. We do not know whether Bob indeed can cheat this much. 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 $|P_L|$?*

For a family L of lines let $\mathbf{X}_L$ denote a random variable in P_L that is a random point on a random line in L. That is, to generate an outcome of $\mathbf{X}_L$, we pick a random line ℓ in L (all lines are equiprobable) and then pick a random point on ℓ (all points on ℓ are equiprobable).

Call any set of lines L satisfying the conditions of Kakeya problem a Kakeya family and let $H(m,q)$ stand for the minimum $H(\mathbf{X}_L)$ over all Kakeya families L. Let $H_\infty(m,q)$ stand for the similar value for min-entropy.

Lemma 9. *Assume that Alice is honest. Then the outcome of the protocol always satisfies $H(y) \geq H(m,q)$ and there is Bob's strategy such that $H(y) = H(m,q)$. The same is true for min-entropy in place of entropy.*

Proof. Let $\mathbf{Y}_S$ stand for the outcome of the protocol provided Bob uses a deterministic strategy S. There is an onto function $S \mapsto L$ from deterministic Bob's strategies to Kakeya sets such that $\mathbf{X}_L$ coincides with $\mathbf{Y}_S$.

Indeed, assume that Bob uses a deterministic strategy S. That is, for each $d = (1, d_2, \ldots, d_m)$ Bob chooses $x = x(d)$ deterministically. Thus Bob defines a Kakeya family L consisting of all lines of the form

$$\{x(d) + td \mid t \in F\}.$$

Obviously $\mathbf{X}_L = \mathbf{Y}_S$.

Conversely, for every Kakeya set L there is Bob's strategy S mapped by this function to L (choose any point in the line in L going in direction d specified by Alice).

This implies the statement of the lemma for deterministic strategies. For randomized strategies it follows from Lemma 2.

Note that for every family of lines L we have $H(Y_L) \leq \log|P_L|$. Thus to prove that the entropy of the outcome is at least α (provide Alice is honest) we need to show the lower bound $|P_L| \geq 2^\alpha$ for Kakeya problem. The best known lower bound for $|P_L|$ is $\Omega(q^{m/2+1})$ [6, 5] (and it is conjectured that $|P_L|$ must be close to q^m). Note that this bound does not immediately imply that $H(\mathbf{Y}_L) \geq (m/2+1)\log q$ for every Kakeya set L, as the entropy of a random variable can be much less than the log-cardinality of the set of outcomes. However, the key proposition from the proof of the bound $|P_L| = \Omega(q^{m/2+1})$ presented in [5] indeed allows to prove a slightly weaker inequality $H(\mathbf{Y}_L) \geq (m/2+1)\log q - O(1)$.

Proposition 1 ([5]). *Let L be a collection of affine lines in F^m such that every 2-dimensional plane has at most $q+1$ lines from L. Let P be a subset of F^m. Then*

$$|\{(p,l) \mid l \in L,\ p \in P,\ p \in l\}| \leq C \cdot (|P|^{1/2}|L|^{3/4}|F|^{1/4} + |P| + |L|)$$

for some constant C.

This proposition allows to prove the following

Theorem 2. *If Alice is honest then the outcome of the geometric protocol satisfies $H(y) \geq (m/2+1)\log q - O(1)$ and $H_\infty(y) \geq \log q$ provided that $m \log q/q \in O(1)$.*

Proof. The second statement is obvious. Let us prove the first one. By Lemma 9 it suffices to show that $H(\mathbf{X}_L) \geq (m/2+1)\log q - O(1)$ for every Kakeya family L.

Let α stand for $q^{-m/2-1}c$ where $c \geq 1$ is a constant to be defined later. We will show that $H(\mathbf{X}_L) \geq -\log\alpha - O(1)$. For each $y \in P_L$ let p_y stand for the probability that $\mathbf{X}_L = y$: that is, p_y is equal to the number of lines in L containing y divided by q^m. We classify y's according to the value of p_y as follows.

Let Q denote the set of those $y \in P_L$ with

$$p_y \leq \alpha$$

and S_i for $i = 1, 2, \dots, -\log\alpha$ the set of those $y \in P_L$ with

$$\alpha 2^{i-1} < p_y \leq \alpha 2^i.$$

The entropy of $\mathbf{X}_L$ is the average value of $-\log p_y$. For all y in Q we have $-\log p_y \geq -\log\alpha$. For all y in S_i we have $-\log p_y \geq -\log\alpha - i$. Thus $H(\mathbf{X}_L)$ can be lower bounded by

$$H(\mathbf{X}_L) \geq -\log\alpha - \sum_i i \cdot \Big(\sum_{y \in S_i} p_y\Big) \geq -\log\alpha - \sum_i i \cdot |S_i| \cdot \alpha 2^i.$$

Thus we need to show that

$$\sum_i i \cdot |S_i| \cdot \alpha 2^i = O(1). \tag{7}$$

To this end we need to upper bound $|S_i|$. We are able to show the following bound.

Lemma 10. *For all i we have $|S_i| \cdot \alpha 2^i = O(2^{-i})$ or $|S_i| \cdot \alpha 2^i = O(q^{-1})$.*

As the series $\sum_i i2^{-i}$ converges and $q^{-1} \cdot (-\log \alpha) \in O(1)$ these bounds obviously imply (7).

Proof. Note that every 2-dimensional plane has at most $q+1$ lines from L (the number of different directions in every plane is equal to $q+1$). Apply Lemma 1 to L and $P = S_i$. We obtain

$$\begin{aligned}|\{(p,l) \mid l \in L,\ p \in S_i,\ p \in l\}| &\le C \cdot (|S_i|^{1/2}|L|^{3/4}q^{1/4} + |S_i| + |L|)\\ &= C \cdot (|S_i|^{1/2}q^{(3m-2)/4} + |S_i| + q^{m-1}).\end{aligned}$$

Every point in S_i belongs more than $\alpha 2^{i-1}q^m$ lines in L hence

$$|\{(p,l) \mid l \in L,\ p \in S_i,\ p \in l\}| > |S_i|\alpha 2^{i-1}q^m.$$

Combining the inequalities we obtain

$$|S_i|\alpha 2^i q^m < C \cdot (|S_i|^{1/2}q^{(3m-2)/4} + |S_i| + q^{m-1}).$$

If the last term in the right hand side is greater than the other ones, we have

$$|S_i|\alpha 2^i < 3C \cdot q^{-1}.$$

If the second term in the right hand side is greater than the other ones, we have

$$\alpha 2^i q^m < 3C.$$

Note that, since $m \ge 2$, $i \ge 1$, we have $\alpha 2^i q^m = 2^i c q^{m/2-1} \ge 2c$. Therefore this cannot be the case, if we let $c \ge 1.5C$.

In the remaining case (the first term in the right hand side is greater than the other ones) we have

$$|S_i|^{1/2} < 3C2^{-i}\alpha^{-1}q^{-m/4-1/2} \Rightarrow |S_i| < 9C^2 2^{-2i}\alpha^{-2}q^{-m/2-1},$$

and

$$|S_i|\alpha 2^i < 9C^2 2^{-i}\alpha^{-1}q^{-m/2-1} = 9C^2 2^{-i}.$$

The last equality holds by the choice of α. Therefore,

$$|S_i|\alpha 2^i \le 9C^2 2^{-i}.$$

If we choose $m = 4$ then the lower bounds for $H(y)$ in the cases when Alice cheats and Bob cheats coincide and are equal to $3\log q - O(1)$. Thus we get:

Theorem 3. *There is a $(3n/4 - O(1), 3n/4 - O(1))$-good 3-round protocol that communicates $2n$ bits.*

Using averaging we obtain the following corollary:

Theorem 4. *There is a $(3n/4 - O(1), 3n/4 - O(1))$-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.

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. O. Goldreich, S. Goldwasser, and N. Linial, Fault-tolerant computation in the full information model. Fault-tolerant SIAM Journ. on Computing 27 (2), 1998.
4. 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.
5. Gerd Mockenhaupt, Terence Tao, Restriction and Kakeya phenomena for finite fields. Duke Math. J. 121 (2004), 35-74.
6. Thomas Wolff, Recent work connected with the Kakeya problem, in Prospects In Mathematics, H. Rossi, ed., AMS 1999.
7. 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.
8. 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.

Appendix

Proof of Lemma. prop-prel Fix c. For $x \in \{0,1\}^*$ let $p_x = \Pr[\mathbf{X} = x]$. For all integer $i \leq n$ let N_i stand for the number of x with

$$2^{-n+i-1} < p_x \leq 2^{-n+i} \tag{8}$$

and w_i for their total probability.

The statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ is equal to

$$\sum_{x:p_x>2^{-n}} (p_x - 2^{-n}) = \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} N_i 2^{-n} \leq \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} 2^{-i} w_i.$$

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

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

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

provided $H(\mathbf{X}) \geq n - c$.

Assume that the entropy of $\mathbf{X}$ is at least $n - c$. The contribution $-p_x \log p_x$ to the entropy of $\mathbf{X}$ of each x satisfying (8) is

$$-p_x \log p_x < -p_x \log 2^{-n+i-1} = p_x(n + 1 - i).$$

Therefore we can estimate the entropy of $\mathbf{X}$ as

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

hence

$$\sum_{i \leq n} i w_i \leq c + 1 \tag{9}$$

Here i ranges over all integers $i \leq n$, including negative ones. However, the contribution of negative i 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

$$0 \geq \sum_{i<0} i w_i \geq \sum_{i<0} i 2^i = O(1).$$

Thus, inequality (9) implies that the sum of $i w_i$ over positive i is bounded by a constant:

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

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:

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

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

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 Appendix), 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 $H(\mathbf{X}) = n - c$. We conjecture that for large enough c the lemma holds for $q = 1 - 2^{-c}$.

A converse of the lemma appears in [7], 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 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 11. *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 $1 - 2^{-c} + e^{-2^c}$.*

Note that the bound of the lemma is almost tight. 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 $H(\mathbf{X}) = n - c$.

Taking a derivative it is easy to verify that $2^c \geq 1 + c \ln 2$ and thus

$$1 - 2^{-c} + e^{-2^c} \leq 1 - (1 - 1/e)2^{-c}$$

for all 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,

$$\sum_z w_z \leq 1. \tag{10}$$

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

$$\sum_z w_z 2^{n-z} \leq 2^n \Rightarrow \sum_z 2^{-z} w_z \leq 1. \tag{11}$$

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

$$\sum_z z w_z \leq c. \tag{12}$$

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 (10), (11) and (12). 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 (11) and (12) imply

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

Note that if $w_z = 0$ for all $z \neq d$ and $w_d = 1$, then (10), (11) and (12) 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

$$\sum_{z>0} w_z (1 - 2^{-z}) \to \max \qquad \text{subject to (10), (11) and (9).}$$

We apply a dual argument. Multiply inequalities (10), (11) and (12) 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

$$0\le\alpha+\frac{\gamma}{\ln 2}+\gamma\log\frac{\beta\ln 2}{\gamma} \tag{13}$$

and

$$1\le\alpha+\frac{\gamma}{\ln 2}+\gamma\log\frac{(1+\beta)\ln 2}{\gamma}. \tag{14}$$

For given β,γ the best α is the minimal one satisfying (13) and (14). It is worth to choose β and γ so that the minimal α satisfying (13) coincides with the minimal α satisfying (14). 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 (13) and (14) 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 $\alpha \geq 0$, which is straightforward.

Remark on the proof. Numerical experiments show that if we let $\beta = 1/(2^{1/\gamma} - 1)$ then for all $c \geq 3.5...$ there is γ such that $\alpha + \beta + c\gamma = 1 - 2^{-d}$. Thus it seems very plausible that for such c the right upper bound for the statistical distance between $\mathbf{U}_n$ and $\mathbf{X}$ with $H(\mathbf{X}) \geq n - c$ is $1 - 2^{-c}$.

Proof of Lemma 2. The first statement is a direct corollary of the properties of conditional entropy. Indeed, let $\mathbf{S}$ be a randomized Bob's strategy and let $\mathbf{Y}_S$ stand for the outcome of the protocol provided Bob uses a deterministic strategy S. Then we have

$$H(\mathbf{Y_S}) \geq H(\mathbf{Y_S}|\mathbf{S}) = \sum_S \Pr[\mathbf{S} = S] H(\mathbf{Y}_S) \geq \sum_S \Pr[\mathbf{S} = S] \cdot \alpha = \alpha.$$

The second statement: we are given that $\Pr[\mathbf{Y}_S = x] \leq 2^{-\alpha}$ for every string x and for every deterministic strategy S. Then for every randomized Bob's strategy $\mathbf{S}$ we have

$$\Pr[\mathbf{Y_S} = x] = \sum_S \Pr[\mathbf{S} = x] \Pr[\mathbf{Y}_S = x] \leq \sum_S \Pr[\mathbf{S} = x] 2^{-\alpha} = 2^{-\alpha}.$$

□

Proof of Lemma 3. 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 $S_i = \{x \in \{0,1\}^n, 2^{-n+i} < p_x \leq 2^{-n+i+1}\}$. It is straightforward that $H(\mathbf{X}) = -\sum_x p_x \log p_x \geq \sum_{i<n} \sum_{x \in S_i} p_x(n - i - 1)$. Since P is resilient and the total probability sums to one, for $0 < i < n$, $\sum_{x \in S_i} p_x < O(2^{c(n-i)}/2^{cn}) = O(2^{-ci})$. Hence, $H(\mathbf{X}) \geq n - 2 - \sum_{0<i<n} i2^{-ci} \geq n - O(1)$. We leave the second part of the claim as an exercise to the interested reader. □

Proof of Lemma 6. We will select a function f satisfying certain properties and we will show that any such function satisfies the lemma and that there are many functions with such properties. Let $K = \{0,1\}^n$ and $L = \{0,1\}^{8 \log n}$. The properties of $f : K \times K \times L \to K$ are as follows (by $f(x, K, L)$ and $f(x, y, L)$ we mean multisets $\{f(x,y,z);\ y \in K\ \&\ z \in L\}$ and $\{f(x,y,z);\ z \in L\}$, resp., and in e.g. $|S \cap f(x, K, L)|$ we count multiplicity):

1. For any $S \subseteq \{0,1\}^n$, where $|S| = 2^n/m$ for $4 < m \leq n^3$,

$$\Pr_{x \in K}[\exists y, |S \cap f(x,y,L)| > \frac{2n^8}{m}] \leq \frac{1}{n^2}.$$

2. For any $S \subseteq \{0,1\}^n$ of size at most $2^n/n^3$,

$$\Pr_{x \in K}[\exists y, |S \cap f(x,y,L)| > 2n^6] \leq \frac{1}{n^2}.$$

3. For any $S \subseteq \{0,1\}^n$ of size at most $2^n/n^{10}$, for all $x \in K$, $|S \cap f(x,K,L)| \leq 2^{n+1}/n^2$.
4. For every $s \in \{0,1\}^n$ and any $x \in K$, $\Pr_{y\in K}[s \in f(x,y,L)] \leq 2n^8/2^n$.

The first two conditions are used to bound the entropy in the case when Alice follows the protocol and Bob may be cheating. The latter two conditions are used to bound the entropy and the min-entropy when Bob follows the protocol. We show that for n large enough the probability that a random function satisfies each of the properties is at least $7/8$, hence a random function satisfies all of them with probability at least $1/2$.

Let $S \subseteq \{0,1\}^n$ be of size $2^n/m$, where $4 < m \leq n^3$. Fix $x, y \in K$. By Chernoff bound (Col A.1.14 of [1]), $\Pr_f[|S \cap f(x,y,L)| > 2n^8/m] \leq e^{-n^8/4m} \leq e^{-n^5/4}$. Hence, for fixed $x \in K$, $\Pr_f[\exists y, |S \cap f(x,y,L)| > 2n^8/m] \leq 2^n e^{-n^5/4} < e^{-n^4}$. Say that $x \in K$ is *bad w.r.t. to* S for honest Alice, if $\exists y \in K$, $|S \cap f(x,y,L)|$ is more than twice its expectation. From the above bound, the probability for a random function f that there are more than $2^n/n^2$ bad x w.r.t. S for Alice is at most $2^{2^n} \cdot \left(e^{-n^4}\right)^{2^n/n^2} < e^{-2^n}$. By union bound over S, with probability at least $1 - 2^{2^n} \cdot e^{-2^n} > 7/8$, for a random function f, for every S of size between $2^n/n^3$ and $2^n/4$, $\Pr_{x\in K}[\exists y, |S \cap f(x,y,L)| > \frac{2n^8}{m}] \leq \frac{1}{n^2}$.

Let $S \subseteq \{0,1\}^n$ be of size at most $2^n/n^3$. Fix $x, y \in K$. By Chernoff bound (Thm A.1.4 of [1]), $\Pr_f[|S \cap f(x,y,L)| > 2n^6] \leq e^{-2n^{12}/n^8} = e^{-2n^4}$. For fixed x, $\Pr_f[\exists y, |S \cap f(x,y,L)| > 2n^6] \leq 2^n \cdot e^{-2n^4} < e^{-n^4}$. Since this happens with very small probability for every x, for a random f there are going to be rather few x's for which there is y such that $|S \cap f(x,y,L)| > 2n^6$. Indeed, by union bound, $\Pr_f[\Pr_x[\exists y, |S \cap f(x,y,L)| > 2n^6] > 1/n^2] \leq 2^{2^n} \cdot e^{-n^4 \cdot 2^n/n^2} < e^{-n2^n} < 1/8$.

For the third condition, let $S \subseteq \{0,1\}^n$ be of size at most $2^n/n^{10}$. Fix $x \in K$. By Chernoff bound, $\Pr_f[|S \cap f(x,K,L)| > 2 \cdot 2^n/n^2] \leq e^{-2\frac{2^{2n}}{n^{12}2^n}} = e^{-2^{n+1}/n^{12}}$. By union bound, $\Pr_f[\exists x, |S \cap f(x,K,L)| > 2 \cdot 2^n/n^2] \leq 2^n \cdot e^{-2^{n+1}/n^12} < 1/8$.

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^8. Thus by Chernoff bound, $\Pr_f[$ the number of occurrences of s in $f(x,K,L) > 2n^8] \leq e^{-n^8/4}$. By a union bound over x and s, with a probability at least $1 - 2^{2n} \cdot e^{-n^8/4} > 7/8$ for a random f, the number of occurrences of any particular element in $f(x,K,L)$ is bounded by $2n^8$, for each x. The property follows.

We claim that if f satisfies the first two properties then the outcome $\mathbf{R}$ of $P_0(Alice, Bob, f)$ has entropy at least $n - O(1)$, assuming that Alice is honest and Bob behaves arbitrarily. For $r \in K$, let $p_r = \Pr[\mathbf{R} = r]$. For $i = 1, \ldots, n-1$, define $S_i = \{r \in K, 2^{-n+r} < p_r \leq 2^{-n+r+1}\}$ and $p_i = \sum_{r \in S_i} p_r$. Clearly, $H(\mathbf{R}) \geq \sum_{i=1}^{n-1} p_i(n-i-1) + (1 - \sum_{i=1}^{n-1} p_i)n \geq n - 1 - \sum_{i=1}^{n-1} p_i i$.

We claim that $\sum_{i=1}^{n-1} p_i i$ is bounded by a constant. As the overall probability is 1, $|S_i| \leq 2^{n-i}$. For $n = 4, \ldots, \log n^3$, by the first property of f, with probability at least $1 - 1/n^2$ Alice picks a random $x \in K$ such that for all y, $|S_i \cap f(x,y,L)| \leq 2n^8/2^i$. Assuming such an x, regardless of which y is chosen by Bob, the output string will fall in S_i with probability at most $2/2^i$. Thus, $p_i \leq 2/2^i + 1/n^2$. For $i = \log n^3, \ldots, n-1$, since $p_i \le 2^{n-i} \le 2^n/n^3$, by the second property of f, with probability $1-1/n^2$ Alice selects a string x so that for all y, $|S_i \cap f(x,y,L)| \le 2n^6$. Thus, $p_i \le 3/n^2$. Hence, $\sum_{i=1}^{n-1} p_i i = O(1)$. This proves the first two parts of the lemma.

For the last part, let $\mathbf{R}$ be the output distribution of the protocol when Bob follows it. The bound on min-entropy follows immediately from the last property of the function f. So we bound the entropy from below. Using the notation from above, for $i = 10\log n, \ldots, n-1$, for all $x \in K$, $|S_i \cap f(x,K,L)| \le 2^{n+1}/n^2$, so $p_i < 1/n^2$. Hence, if y is chosen at random then the probability that there is j such that $f(x,y,j)$ is in S is at most $2/n^2$. The lemma follows by noting: $H(\mathbf{R}) \ge \sum_{i=10\log n}^{n-1} p_i(n-i-1) + (1-\sum_{i=10\log n}^{n-1} p_i)(n-10\log n) > n-3-10\log n$. □

Proof of Lemma 8. WLOG we may assume that Alice uses a deterministic strategy. That is, she chooses a fixed direction d and then choses a fixed point on a random line going in that direction. As different lines in the same direction are pair wise disjoint, all possible q^{m-1} outcomes are equiprobable and thus

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

□

Proof of Lemma 7. In the situation when both players follow the protocol, $P_0(A,B,f_{\mathrm{rot}})$ produces fully random string. If Bob follows the protocol then by analysis similar to that in the proof of Lemma 4 the outcome of the protocol has entropy and min-entropy at least $n-\log n$ as the possibility for Alice to cheat is more constrained.

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. We claim that $\sum_{j=1}^n H(x^j \oplus y) \ge n^2/2 - 3n/2$. Fix $i \ne j \in \{1,\ldots,n\}$. We have

$$H(x^i \oplus y, x^j \oplus y) \ge H(x^i \oplus y \oplus x^j \oplus y) = H(x^i \oplus x^j).$$

The transformation $x \mapsto x^i \oplus x^j$ is linear over the field $\{0,1\}$. That is, $x^i \oplus x^j = A_{i,j}x$ for some matrix $A_{i,j} \in \{0,1\}^{n\times n}$ that depends only on i and j. One can easily verify that since n is prime, $A_{i,j}$ has rank $n-1$. Hence, $H(x^i \oplus x^j) = H(A_{i,j}x) \ge H(x) - \log 2 = n-1$. Thus, $H(x^i \oplus y, x^j \oplus y) \ge n-1$, and

$$\sum_{j=1}^{n} H(x^j \oplus y) \ge \sum_{i=1}^{\lfloor n/2 \rfloor} H(x^{2i} \oplus y, x^{2i+1} \oplus y) \ge n^2/2 - 3n/2.$$

The lemma follows. □