Towards a Reversed Newman's Theorem in Interactive Information Complexity

Joshua Brody[1], Harry Buhrman[2,3], Michal Koucký[4], Bruno Loff*[2], and Florian Speelman[2]

[1]Aarhus University
[2]CWI, Amsterdam
[2]University of Amsterdam
[4]Mathematical Institute of Czech Academy of Sciences, Prague

November 21, 2012

Abstract

Newman's theorem states that we can take any public-coin protocol and convert it into one that uses only private randomness with only a little increase in communication complexity. We consider a reversed scenario in the context of information complexity: can we take a protocol that uses private randomness and convert it into one that only uses public randomness while preserving the information revealed to each player about their own input?

We prove that the answer is yes, at least for protocols that use a bounded number of rounds. As an application, we prove new direct sum theorems through the compression of interactive communication in the bounded-round setting. Furthermore, we show that if a Reversed Newman's Theorem can be proven in full generality, then full interactive communication and direct-sum theorems will hold.

1 Introduction

We will prove in Section REF two results of the following general form: we will take a public-coin protocol π that leaks little information, and "compress" it into a protocol π' that uses little communication to perform the same task. This was our motivation for considering the following:

Question 1. *Is it possible to take a protocol*

*Research supported by grant SFRH/BD/43169/2008, given by FCT, Portugal.

2 Preliminaries

We use capital letters to denote distributions, calligraphic letters to denote sets, and lower-case letters to denote elements in the corresponding sets. So typically A is a distribution over the set $\mathcal{A}$, and a, a', *etc* are elements of $\mathcal{A}$.

2.1 Information Theory

For a given probability distribution A over the support $\mathcal{A}$, its entropy is

$$H(A) = \sum_{a \in \mathcal{A}} p_a \log \frac{1}{p_a},$$

where $p_a = \Pr[A = a]$. Given a second distribution B, the conditional entropy $H(A|B)$ equals

$$\mathbb{E}_{b \in B}[H(A|B = b)].$$

When clear from the context we will denote a conditional distribution $A|B = b$ more succinctly by $A|b$.

We let $I(A : B)$ denote the Shannon information between A and B:

$$I(A : B) = H(A) - H(A|B) = H(B) - H(B|A).$$

Fact 1 (Chain rule)**.**

$$I(A_1 \dots A_k : B|C) = I(A_1 : B|C) + \sum_{i=2}^{k} I(A_i : B|C, A_1, \dots, A_{i-1})$$

Fact 2. *Let A be some distribution and $p_a = \Pr[A = a]$. Then*

$$\Pr_{a \in A}[p_a < 2^{-\frac{1}{\varepsilon} H(A)}] \leq \varepsilon.$$

2.2 Communication and Information Complexities

We assume that the reader is familiar with communication complexity. We will be dealing with protocols that have both public and private randomness; this is not very common, so we will give the full definitions, which are essentially those of (REF). We will also be working exclusively in the distributional setting, but all our compression and direct theorem results will follow in the usual setting, with roughly the same parameters, by the use of Yao's Lemma (REF) and its Information Complexity variants (REF Braverman '10). So from here onwards, we will assume that the input is given to two players, Alice and Bob, by sampling from a possibly correlated distribution $\mu = (X, Y)$ over the support $\mathcal{X} \times \mathcal{Y}$.

A *private coin protocol* π is defined as a rooted tree, called the *protocol tree*, in the following way:

1. Each non-leaf node is owned by either Alice or Bob.

2. If v is a non-leaf node belonging to Alice, and it is the j-th such node in its path to the root, then:
   (a) The children of v are owned by Bob and form a set $\mathcal{C}(v) \subseteq \{0,1\}^*$;
   (b) Associated with v is a set $\mathcal{R}_v$, and a function $M_v : \mathcal{X} \times \mathcal{R}_v \to \mathcal{C}(v)$.
3. The situation is analogous for Bob's nodes.
4. With each a leaf we associate an *output value*.

On input x, y the protocol is executed as follows:

1. Set v to be the root of the protocol tree.
2. If v is a leaf, the protocol ends and outputs the value associated with v.
3. If v is owned by Alice, she picks a string r uniformly at random from $\mathcal{R}_v$ and sends $M_v(x, r)$ to Bob, they both set $v = M(x, r)$, and return to the previous step. Bob proceeds analogously on the nodes he owns.

A general protocol is a distribution over private coin protocols. The players run such a protocol by using shared randomness to pick an index r (independently of X and Y) and then executing the corresponding private-coin protocol π_r. A protocol is called *public-coin* if every $\mathcal{R}_v$ has size 1, i.e., no private randomness is used.

Definition 1. *The* communication complexity *of a protocol π, $\mathsf{CC}_\mu(\pi)$, is the average number of bits that are transmitted in an execution of π on input distribution μ. The* number of rounds *of π, $\mathsf{RC}_\mu(\pi)$, is the average depth reached in the protocol tree by an execution of π on input distribution μ.*

Observation 1. *If $\mathsf{CC}(\pi) = C$, then we can assume that each $\mathcal{R}_v$ is the set of $O(C)$-long binary strings. We leave the proof for the full version of the paper.*

We let $\pi(x, y, r, r^{(a)}, r^{(b)})$ denote the messages exchanged during the execution of π, for given inputs x, y, and random choices $r, r^{(a)}$ and $r^{(b)}$), and $\textsc{Out}_\pi(x, y, r, r^{(a)}, r^{(b)})$ be the output of π for said execution. We use R to denote the public randomness distribution, $R^{(a)}$ for Alice's randomness, and $R^{(b)}$ for Bob's randomness; we use Π to denote the distribution $\pi(X, Y, R, R^{(a)}, R^{(b)})$.

Definition 2 (REF barak et al braverman, ...?)**.** *The* (internal) information cost *of protocol π with respect to μ is:*

$$\mathsf{IC}_\mu(\pi) = I(Y : R, \Pi | X) + I(X : R, \Pi | Y).$$

Definition 3. *A protocol π is said to compute function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ with error ε over distribution μ if*

$$\Pr_{\mu, R, R^{(a)}, R^{(b)}}[\mathrm{Out}_\pi(x, y, r, r^{(a)}, r^{(b)}) = f(x, y)] \geq 1 - \varepsilon.$$

We will be particularly interested in the case of one-way protocols, and hence we will define this special case with its own nomenclature and notation:

Definition 4. *A* one-way private-coin protocol *is given by a function* $M : \mathcal{X} \times \mathcal{R} \mapsto \mathcal{M}$. *On input* $x \in \mathcal{X}$, *Alice uniformly picks a random* $r \in \mathcal{R}$ *using her own private randomness, and sends* $M(x, r)$ *to Bob.*

The protocol is said to be 1–1 if $M(x, \cdot)$ *is 1–1 for all* x.

The information cost *of a one-way private-coin protocol, with respect to some distribution* X *on* $\mathcal{X}$, *is the Shannon information* $I(X : M(X, R))$.

In a one-way public-coin *protocol Alice proceeds the same way, except that now* r *is picked using shared randomness.*

Since both players know R, *the information cost of a one-way public-coin protocol is instead* $I(X : M(X, R), R)$.

3 Compression for public-coin protocols

We will prove in this section two results of the following general form: we will take a public-coin protocol π that leaks little information, and "compress" it into a protocol ρ that uses little communication to perform the same task with about the same error-probability. It turns out that the results in this setting are simpler and give stronger compression than in the case where Alice and Bob have private randomness (such as in REF). We will prove two bounds, one that is dependent on the number of rounds of π, but which is also round-efficient, in the sense that ρ will not use many more rounds than π; and one that is independent of the number of rounds of π, but where the compression is not as good when the number of rounds of π is small.

If the number of rounds is high, and we do not care about the number of rounds of our compression protocol but only about its communication complexity, then we can get a very strong compression result for public coin protocols (compare with the best known compression for private-coin protocols, that achieves $\tilde{O}(\sqrt{IC})$ communication):[1]

Theorem 3. *Suppose there exists a public-coin protocol* π *to compute* f *over the input distribution* $\mu = (X, Y)$ *with error probability* ε, *and let* $C = \mathsf{CC}(\pi)$, $I = \mathsf{IC}_\mu(\pi)$. *Then there exists a public-coin protocol* ρ *that computes* f *over* μ *with error* $\varepsilon + \delta$, *and with*

$$\mathsf{CC}(\rho) \leq O\left(\delta^{-3} I \log C)\right).$$

Let us sketch the proof of Theorem 3. If the information cost $\mathsf{IC}_\mu(\pi)$ is small, then from Bob's perspective it holds that:

$$I(X : \Pi | R, Y) = H(\Pi | R, Y) - H(\Pi | R, X, Y) \leq I.$$

[1] It appears that Theorem 3 was discovered independently by Denis Pankratov (REF). Our proof is a bit different: we discovered it originally while studying the compression problem in a Kolmogorov complexity setting (as in REF harry michal), and our proof for the Shannon setting arises from the proper "translation" of that earlier proof, whereas Pankarov's proof is in a style similar to Barak et al. (REF barak et al).

But $H(\Pi|R,X,Y)=0$, because since π is a public-coin protocol then the transcript Π is a function of R,X,Y. Hence from Bob's perspective the entropy of Π, $H(\Pi|R,Y)$, is at most I, which means that with high probability it will hold that $H(\Pi|r,y)=h_{r,y}\leq O(I)$. Now given the public randomness r and his input y, Bob knows the distribution $\Pi|r,y$, and he can hence form the set of transcripts

$$\Pi^{(b)}=\{\tau|\Pr[\Pi=\tau|r,y]\geq 2^{-O(I)}\}.$$

It will hold from Fact 2 that the true transcript $\pi(x,y,r)$ will be in $\Pi^{(b)}$ with high probability; nonetheless, it also holds that $\Pi^{(b)}$ has only $2^{O(I)}$ many transcripts. Alice can do the same thing, and come up with a small set $\Pi^{(a)}$ of transcripts, and it will hold w.h.p. that $\Pi^{(a)}\cap\Pi^{(b)}=\{\pi(x,y,r)\}$. Now all that is necessary is a procedure to find the unique intersection of the two sets $\Pi^{(a)}$ and $\Pi^{(b)}$, since $\pi(x,y,r)$ is enough to compute $f(x,y)$, or for that matter perform any task that can be accomplished through π. In general finding the intersection of two sets of such size would require $2^{O(I)}$ bits of communication (REF), but we will see that since the elements of $\Pi^{(a)}$ and $\Pi^{(b)}$ are transcripts, Alice and Bob will have enough information to make the search exponentially more efficient.[2] The full proof is left for the appendix.

It is possible to compress a public-coin protocol on a round-by-round basis while preserving, up to a multiplicative constant, the total number of rounds used. This is in contrast to the compression scheme of (REF braverman amortized), which also compresses one round at a time, but where the compressed version of the protocol pays a super-constant multiplicative increase in the number of rounds.

Theorem 4. *Suppose there exists a public-coin, q-round protocol π to compute f over the input distribution $\mu=(X,Y)$ with error probability ε, and let $C=\mathsf{CC}(\pi)$, $I=\mathsf{IC}_\mu(\pi)$. Then there exists a public-coin, $O(\delta^{-1}q)$-round protocol ρ that computes f over μ with error $\varepsilon+\delta$, and with*

$$\mathsf{CC}(\rho)\leq O(\delta^{-1}(I+q\log q)).$$

The idea of the proof is to show the result one round at a time. In round i, Alice, say, must send a certain message m_i to Bob. From Bob's point of view, this message is drawn from some distribution $M_i=M_i(\tilde{X},y,r,m_1,\ldots,m_{i-1})$ where $\tilde{X}$ is Alice's input distribution conditioned on Bob's input y, on the public randomness r, and on the messages $m_1,\ldots,m_{i-1}$ that were previously exchanged. However, we will show that there is a sub-protocol σ_i that can simulate round i with very small error by using constantly-many rounds and

$$O(H(M_i|y,r,m_1,\ldots,m_{i-1}))=I(X:M_i|y,r,m_1,\ldots,m_{i-1})$$

bits of communication *on average*. Then putting these sub-protocols together, and truncating the resulting protocol whenever the communication or the number of rounds is excessive, we obtain the protocol ρ which simulates π.

[2]The basic property required is that given any prefix of the string in the intersection of $\Pi^{(a)}$ and $\Pi^{(b)}$, either Alice or Bob know how to correctly extend it by one bit.

The procedure to compress each round is achieved through an interactive variant of the one-shot Sleepian-Wolf theorem (REF sw, renner, buhrman). We could not apply this theorem directly, however, since it is made to work on a more restrictive, one-way scenario (whereas we are allowed a constant number of rounds on average), and the error guarantee in that case is not enough to put the different sub-protocols together with a sufficiently small enough total error. It should be possible to prove a similar result that works with a single additional round, but we will leave this for the full version of the paper. The full proofs for this section are in the appendix.

3.1 A new strategy for proving compression theorems in the general case where private randomness is involved

We have proven that very good compression is possible if only public randomness is used. This is perhaps to be expected, because of the following intuition: unless private randomness is involved, how are Alice and Bob supposed to "hide" their inputs from each other? This intuition seems to be corroborated by the following simple scenario:

Consider the (toy) protocol where Alice is given an input x drawn according to some distribution X, then she uses her private randomness to pick a uniformly-random string r of the same length as x, and sends Bob the bit-wise parity $x \oplus r$. It is clear that this reveals zero information about X to Bob. However, if Alice were to pick r using public coins instead of private coins, then Bob could easily and always recover x from $x \oplus r$, and hence the protocol would give full information about X ($H(X)$).

So it would seem that the use of private randomness is necessary, and in fact in the application of information complexity to the proof of direct sum theorems, private randomness appears in a natural, seemingly unavoidable way. We will show, however, that this is not the case, that it is possible to take a private-coin protocol π and convert it into a public coin protocol $\tilde{\pi}$ that has the exact same transcript distribution, and not much higher information cost. More precisely, the rest of the paper is mostly devoted towards proving the following theorem:

Theorem 5. *Let π be an arbitrary private-coin, q-round protocol, and let $C = \mathsf{CC}(\pi)$. Suppose that π's public randomness is chosen from some uniform distribution R over the set $\mathcal{R}$, and π's private randomness is chosen from uniform distributions $R^{(a)}$ and $R^{(b)}$ over the sets $\mathcal{R}^{(a)}$ and $\mathcal{R}^{(b)}$.*

Then there exists a public-coin, q-round protocol $\tilde{\pi}$, whose public randomness is chosen from the distribution $R \times R^{(a)} \times R^{(b)}$, and that has the exact same transcript distribution, i.e.,

$$\forall x, y, t \Pr_{r \in R, r^{(a)} \in R^{(a)}, r^{(B)} \in R^{(b)}}[\pi(x, y, r, r^{(a)}, r^{(b)}) = t] = \Pr_{\vec{r} \in R \times R^{(a)} \times R^{(b)}}[\pi(x, y, \vec{r}) = t],$$

and such that for any input distribution $\mu = (X, Y)$,

$$\mathsf{IC}_\mu(\tilde{\pi}) \leq \mathsf{IC}_\mu(\pi) + O(q \log(n + C)).$$

The theorem seems very counter-intuitive, in addition to claiming something which we intuitively think shouldn't be possible, there seems to be hardly any leg-room for changing π at all. Actually, there is one change that we are allowed to make. If Alice, for instance, wishes to send a message $M = M(x, r^{(a)})$ in protocol π, and noticing that $r^{(a)}$ is picked uniformly, she might instead send message $M(x, \phi_x(r^{(a)}))$, where ϕ_x is some permutation of $\mathcal{R}^{(a)}$. The permutation ϕ_x will somehow "scramble" the formerly-private now-public randomness $R^{(a)}$ into some new string $\tilde{r}^{(a)} = \phi_x(r^{(a)})$ about which Bob hopefully knows nothing. This "scrambling" keeps the protocol exactly as it was, changing only the association saying which private randomness results in each message. We will see that this can be done in such a way that, in spite of knowing $r^{(a)}$, Bob has no hope of knowing $\tilde{r}^{(a)} = \phi_x(r^{(a)})$, unless he already knows x to begin with.

From the combination of Theorems 4 and 5, we can obtain a new compression result for general protocols.

Corollary 6. *Suppose there exists a public-coin, q-round protocol π to compute f over the input distribution $\mu = (X, Y)$ with error probability ε, and let $C = \mathsf{CC}(\pi)$, $I = \mathsf{IC}_\mu(\pi)$. Then there exists a public-coin, $O(\delta^{-1}q)$-round protocol ρ that computes f over μ with error $\varepsilon + \delta$, and with*

$$\mathsf{CC}(\rho) \leq O(\delta^{-1}(I + q \log(q(n + C)))).$$

As we will see in the following sub-section, this will result in a new direct sum theorem for bounded-round protocols, which is a slight improvement over the results of (REF braverman amortized). In general, given that we have already proven Theorem 3, and given that this approach shows promise in the bounded-round case, it becomes worthwhile to investigate wether we can make the randomness public while leaking little information, in any general protocol with an unbounded number of rounds. Hence the title of our paper.

The next theorem will be proven in sections 5 and 4. We will generally call "Reversed Newman's theorem" to those results that make randomness public without leaking more information.

Theorem 7 (Reversed Newman's theorem for one-way protocols)**.** *Given a one-way private-coin protocol M, there is a one-way public-coin protocol M' generating the same message distribution, and such that*

$$I(X : M'(X, R), R) \leq I(X : M(X, R)) + O(\log\log(|\mathcal{X}||\mathcal{R}|)),$$

for any input distribution X.

Let us prove Theorem 5 as a consequence of Theorem 7.

Proof of Theorem 5. The new protocol $\tilde{\pi}$ first picks public randomness r according to R and then proceeds by simulating $\pi|_r$ a round-per-round basis. By Fact (TODO REF) we can assume that the private randomness used in each round is a binary string of length $O(C)$, picked uniformly at random and independently

from the remaining rounds; i.e., that $R^{(a)} = R^{(a,1)} \times \ldots \times R^{(a,q)}$ for uniform distributions $R^{(a,j)}$ over $\{0,1\}^{O(C)}$, and the same for $R^{(b)}$, and that message π_j for round j depends only on x (or y), $r^{(a,j)}$ (or $r^{(b,j)}$), and on the previous messages. Now suppose we wish to simulate round $j+1$ of π, and that we have already successfully simulated π up to round j, which cost us at most

$$I(X : \Pi_{\leq j}|Y, r) + I(X : \Pi_{\leq j}|Y, r) + O(j \log(n + C))$$

bits of information. Suppose w.l.o.g. that it is Alice's turn to communicate, and that messages $m_1, \ldots, m_j$ have been exchanged up to this point. Then round $j+1$ is simulated thus: we consider the one-way private-coin protocol $M(x, r^{(a,j+1)}) = \pi_{j+1}(r, x, r^{(a,j+1)}, m_1, \ldots, m_j)$, and we run M' instead, as given by Theorem 7. As per that theorem, for any given y the information leaked will be

$$I(X : M'(X, R^{(a,j+1)}), R^{(a,j+1)}|r, y, m_{\leq j}) \\ \leq I(X : M(X, R^{(a,j+1)})|r, y, m_{\leq j}) + O(\log(n + C)).$$

When averaged over Y and $\Pi_1, \ldots, \Pi_j$ (recall that the message distribution is always preserved), the left-hand side becomes the information cost of our new protocol on round $j+1$. The right-hand side is $I(X : \Pi_{j+1}|Y, r, \Pi_{\leq j}) + O(\log(n+C))$, which when added to the previous information cost becomes exactly

$$I(X : \Pi_{\leq j+1}|Y, r) + I(X : \Pi_{\leq j+1}|Y, r) + O((j+1) \log(n + C)).$$

□

3.2 Direct sum theorems for bounded rounds

The following theorem was proven in (REF barak et al)

Theorem 8. *(REF theorem number) Suppose that there is a q-round protocol π^k that computes k copies of f with communication complexity C and error ε. Then there exists a q-round private-coin protocol π that computes a single copy of f with communication complexity C and the same error probability ε, but with information cost $\mathsf{IC}_\mu(\pi) \leq \frac{C}{k}$ for any input distribution μ.*[3]

As a consequence of this theorem, and of Corollary 6, we will be able to prove the following:

Theorem 9 (Direct sum theorem for the bounded-round case)**.** *Let $\varepsilon > 0$, $\delta > \varepsilon$ be two constants, then there is some constant $d \geq \delta^{-1}$ such that, for any input distribution μ, if f requires $d(C + q \log q(n + C))$ bits of communication*

[3] In the style of (REF Braverman information complexity 2012), we would say that if the communication complexity of $f^{\otimes k}$ is at most C, then the *information complexity* of f is at most C/k.

to be computed over μ with error δ in dq rounds, then $f^{\otimes k}$ requires kC bits of communication to be computed over $\mu^{\otimes k}$ with error ε in q rounds. In particular:

$$\frac{R^q_\varepsilon(f^{\otimes k})}{k} \geq \Omega\left(R^{dq}_\delta(f) - O\left(q\log\left(qn + qR^{dq}_\delta(f)\right)\right)\right).$$

This theorem follows from a weaker hypothesis than that of (REF braverman amortized). Namely, in that paper, for the same constant-error case, it is proven that

$$\frac{R^q_\varepsilon(f^{\otimes k})}{k} \geq \Omega\left(R_\delta(f) - O(q) - O\left(\sqrt{qR_\delta(f)}\right)\right).$$

Notice that in the above bound the communication complexity of f is with respect an unrestricted number of rounds. Also, in a scenario where more than $(\log n)^2$ bits-per-round are sent, these new bounds are better. Proof is left to the appendix.

4 R.N.T. for 1–1, one-way protocols

Before we are able to prove Theorem 7, let us prove the special case of 1–1 protocols.

Theorem 10. *Given a one-way private-coin 1–1 protocol M, there is a one-way public-coin protocol M' generating the same message distribution, and such that*

$$I(X : M'(X,R),R) \leq I(X : M(X,R)) + O(\log\log|\mathcal{X}|),$$

for any distribution X.

The proof will be left for appendix, and here we will only sketch it. Let us think of protocol M as being given by a table. The rows of this table are the inputs x, and the columns are the choices r of private randomness. For simplicity suppose that Alice's input is uniformly distributed in some set $\mathcal{X}$. Suppose that Bob has received message m; in the scenario when he does not know Alice's private randomness, the uncertainty that remains after receiving m is the entropy $H(X|M(X,R)=m)$ which with our simplifying assumption equals

$$\log|S_m|: \qquad S_m = \{x|\exists r\ M(x,r)=m\}.$$

However, in the situation where Bob *does know* r, then the uncertainty is now

$$\log|S_{m,r}|: \qquad S_{m,r} = \{x|M(x,r)=m\}.$$

It could happen that $S_{m,r}$ has on average fewer elements than S_m, which would result on a smaller uncertainty about x and hence a larger information cost. However, if we could somehow make sure that every set S_m gets broken into at most d different sets $S_{m,r}$, then we intuitively expect the entropy $\log|S_{m,r}|$ to be only $\log d$ bits smaller than $\log|S_m|$.

So our strategy to prove Theorem 10 will be to find a way of permuting each row of our table, in such a way that any message m is split among few columns. This can be approximately achieved by a certain combinatorial construction, which we call a *matching graph*.

Definition 5. *An (m, n, d, δ)-matching graph is a bipartite graph $G = (\mathcal{A} \cup \mathcal{B}, \mathcal{E})$ such that $|\mathcal{A}| = m$, $|\mathcal{B}| = n$, $\deg(u) = d$ for each $u \in \mathcal{A}$, and such that for all $\mathcal{A}' \subseteq \mathcal{A}$ with $|\mathcal{A}'| = n$, $G_{\mathcal{A}' \cup \mathcal{B}}$ has a matching of size at least $n(1-\delta)$.*

In the appendix, we will use the probabilistic method to prove the existence of matching graphs with sufficiently good parameters:

Lemma 11. *A (kN, N, d, δ)-matching graph exists with $d = \frac{2+\ln k}{\delta^2} + \frac{\ln(1/\delta)}{\delta}$.*

Now the proof of Theorem 10 proceeds roughly like this: we let $\mathcal{A} = \mathcal{M}$ be the set of all messages, $\mathcal{B} = \mathcal{R}$ be the set of all random choices, and then, for each row of our table, we let $\mathcal{A}' = M(x, \mathcal{R})$ be the set of messages on that row, and extend the partial matching between $\mathcal{A}'$ and $\mathcal{R}$ to a perfect matching $\phi_x : \mathcal{R} \to \mathcal{A}'$. The new protocol M' sets $M'(x, r) = \phi_x(r)$ for each $r \in \mathcal{R}$. Then it will be seen that, at least approximately, no message m can be sent to more than d different columns, where d is the degree of our graph, and that this is enough to ensure that the uncertainty about x is preserved, even when the randomness is made public. See the appendix for the proof.

5 R.N.T. for general one-way protocols

We will finally prove Theorem 7. Let us begin by proving how Theorem 7 follows from Theorems 10 together with the following:

Theorem 12 (Making the protocol 1–1 while leaking little information)**.** *Given a one-round private-coin protocol M, there is a one-round 1–1 private-coin protocol $\tilde{M}$ of the form*

$$\tilde{M}(x, r) = (M(x, r), J(x, r)),$$

and such that

$$I(X : J(X, R) | M(X, R)) \leq O(\log\log|\mathcal{X}||\mathcal{R}|).$$

To prove Theorem 7, we begin with a one-way protocol M, and let $\tilde{M} = (M, J)$ be given by Theorem 12; $\tilde{M}$ has the property that

$$I(X : \tilde{M}(X, R)) = I(X : M(X, R)) + I(X : J(X, R) | M(X, R)),$$

and $I(X : J(X, R)|M(X, R)) \leq O(\log\log|\mathcal{X}||\mathcal{R}|)$. We then apply Theorem 10 to $\tilde{M}$, to get a protocol $\tilde{M}'$, such that

$$I(X : \tilde{M}'(X, R), R) \quad \leq \quad I(X : \tilde{M}(X, R)) + O(\log\log|\mathcal{X}|).$$

Our protocol M' will be the first part of $\tilde{M}' = (M', J')$, which of course leaks no more information than $\tilde{M}'$, and is distributed exactly like M. QED.

Now let us sketch the proof of Theorem 12. The message $\tilde{M}$ is defined in a very straightforward way. We again think of M as a table, and fix some ordering $r_1 < r_2 < \ldots$ of $\mathcal{R}$. Then the second part $J(x, r)$ will be set to the number of times $M(x, r)$ appeared in the same row up to the column r, i.e.,

$$J(x, r) = |\{r' \leq r | M(x, r') = M(x, r)\}|.$$

Then we will prove that, for every message m, the following holds:

$$I(X : J(X, R)|m) = H(X|m) - H(X|m, J(X, R)) \leq O(\log \log |\mathcal{X}||\mathcal{R}|).$$

How do we interpret the quantities above? Let us give some intuition by considering the simpler case where X is uniformly distributed. Then $X|m$ is the following distribution: for w_x the number of times that m appears in row x, and $T = \sum_x w_x$ be the total number of times m appears in the protocol table, we have $\Pr[X = x|m] = w_x/T$.

Or, we can think of $X|m$ as being given by the following process: we have *shape* $\mathcal{S}$ formed by a collection of squares, arranged in left-justified rows, one row for each x, and w_x squares on row x, denoted $(x, 1), (x, 2), \ldots, (x, w_x)$. Then $X|m$ is the distribution obtained by selecting the row x from a randomly chosen square (x, j) in the shape $\mathcal{S}$. With this second image, it is now easy to see that for a given j, $X|m, j$ is the distribution obtained by selecting the row x from a randomly chosen square (x, j) (j is now fixed).

Hence, we wish to bound the difference between the overall entropy $H(X|m)$ of the distribution induced by the entire shape $\mathcal{S}$, and the average entropy $H(X|m, J)$ of the distributions induced by the columns of $\mathcal{S}$. Obtaining this bound, it turns out, is not trivial. The proofs and required definitions are left to the appendix.

A Proofs of Section 3

A.1 Proof of Theorem 4

Let us begin by showing how we may compress one round. Suppose that Alice and Bob are given inputs drawn according to the distribution $\tilde{\mu} = (\tilde{X}, \tilde{Y})$, and that Alice wishes to send a message $m = M(\tilde{x})$ to Bob which is the result of applying the function M to her input. Suppose that this costs her I bits of information, i.e., that

$$I(\tilde{X} : M(\tilde{X})|\tilde{Y}) = I.$$

Then we will construct a protocol σ that succeeds in communicating message m to Bob, with error probability ε, and using only, on average, a constant number of rounds and $O(\lceil I \rceil + \log \frac{1}{\varepsilon})$ bits of communication. More formally:

Theorem 13. *Let M be any deterministic one-way protocol, and let $\mu = (X, Y)$ be some input distribution. Then there exists a randomized protocol σ, together with a decoding functions d, with the following properties.*

1. *Let T denote a random variable made of the public randomness together with the transcript of a run of σ on inputs drawn according to μ;*
2. *Then there is an event Ω, which we call the "error event," that happens on such a run with probability at most δ;*[4] *and*
3. *Conditioned on $\neg\Omega$, it holds that σ uses an average of $O(1)$ rounds and an average of $O(I + \log \varepsilon^{-1})$ bits of communication; and*
4. *Conditioned on $\neg\Omega$, it holds that $d(t, y) = M(x)$.*

Proof. For a given choice of inputs, let M_y denote the distribution $M(\tilde{X})$ conditioned on the event $\tilde{Y} = y$, let $m = M(x)$ be Alice's message, and let ℓ be an upper bound on the length of $M(\tilde{X})$.

The protocol proceeds as follows. Bob calculates the quantity

$$h_y = I(\tilde{X} : M(\tilde{X})|y) = H(M_y),$$

by computing the probability masses of the distribution M_y. Then Bob sends Alice the value $\lceil h_y \rceil$ in binary, and Alice and Bob begin communicating a certain number of stages, each stage requiring two rounds, to try and let Bob know what m is. In stage 1, Bob begins by considering the set

$$S_1 = \{m' | \Pr[M_y = m'] \geq 2^{-3h_y}\}.$$

Then both players use their shared randomness to pick a random linear function $F_1 : \mathbb{F}_2^\ell \to \mathbb{F}_2^k$, with $k = 3\lceil h_y \rceil + \lceil \log \frac{2}{\varepsilon} \rceil$, and Alice sends m's *fingerprint* $f_1 = F_1(m)$ to Bob.

Bob will look for the first $m' \in S_1$ such that $F_1(m') = f_1 = F_1(m)$ (i.e., having the same fingerprint as m). If such an m' exists, then Bob replies indicating

[4] The probability is taken over μ and over the public randomness of σ.

that the protocol is over and assumes that $m' = m$, i.e., the decoding will set $d(h_y, F_1, f_1, y) = m'$. If no such m' exists, then Bob knows that $m \notin S_1$, and replies indicating that the protocol should continue. On the j-th stage of doing this, Bob looks at the set:

$$S_j = \{m' | \Pr[M_y = m' | M_y \notin S_1 \cup \ldots \cup S_{j-1}] \geq 2^{-3h_y}\}.$$

Then they repeat what they did in the first stage: they pick a shared random F_j and Alice sends $F_j(m)$ to Bob. Bob will look for the first $m' \in S_j$ such that $F_j(m') = f_j = F_j(m)$ and $m' \notin S_1 \cup \ldots \cup S_{j-1}$. If such an m' exists, then Bob replies indicating that the protocol is over and assumes that $m' = m$. If no such m' exists, then Bob assumes that $m \notin S_1 \cup \ldots \cup S_j$, and replies indicating that the protocol should continue.

By construction it is clear that $|S_j| \leq 2^{2h_y}$, and hence it can easily be seen that, for any fixed j, with probability $1 - \frac{\varepsilon}{2}$ there will exist no string $m' \neq m$ in S_j with the same fingerprint as m.

Now let us analise the first stage. Using Fact 2, we find that $m \in S_1$ will happen with probability $\geq 2/3$ over the choice of x. If $m \in S_1$, then it will hold with probability $1 - \frac{\varepsilon}{2}$ that m is the only string in S_1 with the same fingerprint, and hence Bob will find m. Thus, with probability $\geq (1 - \frac{\varepsilon}{2})\frac{2}{3} > 1/2$, σ will be over in a single stage with a correct decoding, and in this case the total number of bits communicated is $\lceil \log h_y \rceil + k + 1$. We say that there is an error during the first stage if $m \notin S_1$ but there is a string $m' \in S_1$ with the same fingerprint, and this happens with probability $\leq \frac{\varepsilon}{2}$.

Now let us analise stage j. Conditioned on the event $E_j \equiv M_y \notin S_1 \cup \ldots \cup S_{j-1}$, the entropy of M_y can only go down and is hence at most h_y, so we find that we may use Fact 2 again, to conclude that, conditioned on E_j, $m \in S_j$ will happen with probability $\geq 2/3$ over the choice of x. Hence, once again, with probability $> 1/2$, σ will conclude at the end of stage j with the correct message; and in this case the total communication is $\lceil \log h_y \rceil + j(k+1)$. An error at stage j happens if we reach stage j with $m \notin S_j$, but there is a string $m' \in S_j$ with the same fingerprint as m; conditioned on having reached stage j, an error will happen with probability $\leq \frac{\varepsilon}{2}$, and also in this case the protocol concludes at the end of stage j, but now possibly with an incorrect message.

It can thus be seen that the protocol reaches stage j with probability at most 2^{-j+1}. Hence the average number of stages is at most

$$1 + \frac{1}{2}2 + \frac{1}{4}3 + \ldots = 1 + \sum_{i=1}^{\infty} \frac{i+1}{2^i} \leq 5,$$

and the average communication is at most

$$\lceil \log h_y \rceil + (k+1) + \frac{1}{2}(k+1) + \frac{1}{4}2(k+1) + \ldots \leq \lceil \log h_y \rceil + 3(k+1) \leq 5\lceil h_y \rceil + 3\log\frac{2}{\varepsilon} + 2.$$

Now if we let Ω denote the event of having an error at any stage, then conditioned on the non-occurrence of Ω Bob will know the correct message m, and

furthermore the probability that Ω occurs will be at most

$$\frac{\varepsilon}{2} + \frac{1}{2}\frac{\varepsilon}{2} + \frac{1}{4}\frac{\varepsilon}{2} + \ldots \leq \varepsilon.$$

□

Now we proceed to show compression for a general protocol π with q rounds. We may assume that π is deterministic, since for a public coin protocol π, its information cost $\mathsf{IC}_\mu(\pi)$ is simply the average over the shared randomness r of $\mathsf{IC}_\mu(\pi_r)$ (where π_r is the protocol that runs π with randomness r); hence if we can compress each π_r, then we can compress π. We may further suppose without loss of generality that π behaves in a way such that Alice communicates in the odd-numbered rounds and Bob communicates in the even-numbered rounds, and furthermore the message sent in round j is always of length exactly $\ell(j)$.[5] With this supposition it is now clear that:

$$\mathsf{IC}_\mu(\pi) = \sum_{\text{odd } j \in [q]} I(X : M_j | Y, \vec{M}_{<j}) + \sum_{\text{even } j \in [q]} I(Y : M_j | X, \vec{M}_{<j})$$

where M_j is the distribution of the j-th message of π. Let us denote the information cost of round j with I_j.

We begin by proving an average-case variant of Theorem 4. Informally, the theorem says that for any q-round protocol π there exists a randomized protocol $\tilde{\rho}$ that simulates π with error δ, while using $O(q)$ rounds and $O(I + q \log \frac{q}{\delta})$ bits of communication *on average*. This is written more precisely as follows.

Theorem 14. *Let π be any deterministic q-round protocol, and let $\mu = (X, Y)$ be some input distribution. Then there exists a randomized protocol $\tilde{\rho}$, together with two decoding functions $d^{(a)}$ and $d^{(b)}$, with the following properties.*

1. *Let T denote a random variable made of the public randomness together with the transcript of a run of $\tilde{\rho}$ on inputs drawn according to μ.*
2. *Then there is an event Ω, which we call the "error event," that happens on such a run with probability at most δ; and*
3. *Conditioned on $\neg\Omega$, it holds that $\tilde{\rho}$ uses an average of $O(q)$ rounds and an average of $O(I + q \log(\delta^{-1} q))$ bits of communication; and*
4. *Conditioned on $\neg\Omega$, it holds that $d^{(a)}(t, x) = d^{(b)}(t, y) = \pi(x, y)$.*

Proof. Let π be a given deterministic q-round protocol. We define a protocol $\tilde{\rho}$ that simulates π by using the one-round compression of Theorem 13 on a round-per-round basis.

The first round, for instance, is simulated by a run of the protocol σ of Theorem 13 for the one-way protocol that is sending the first message of π, i.e.,

[5] This is done by filling the extra message bits with zeros. This causes the number of rounds to be at most doubled, and the communication complexity to be multiplied by at most q, while ensuring that the information cost remains exactly the same.

$M = M_1$, $\tilde{X} = X$ and $\tilde{Y} = Y$. The error parameter is set to $\varepsilon = \frac{\delta}{q}$, and after the execution of σ Bob will know some message $\tilde{m}_1$ which may or may not be equal to $M_1(x)$. Let Ω_1 denote the event that $\tilde{m}_1 \neq M_1(x)$. Then with probability $\Pr[\neg\Omega_1] \geq 1 - \varepsilon$, $\tilde{m}_1 = M_1(x)$, and furthermore, conditioned on $\neg\Omega_1$, we can see that, on average, σ only used a constant number of rounds and $O(I_1 + \log \frac{q}{\delta})$ bits of communication.

Now suppose that j rounds have been simulated correctly ($\neg\Omega_{\leq j}$), that it is Alice's turn again ($j+1$ is odd), and that both players communicated (the encoded version of) the messages $\tilde{m}_1, \ldots, \tilde{m}_j$; let $\tilde{X}$ and $\tilde{Y}$ be Alice and Bob's input distributions conditioned on these events; i.e.,

$$\tilde{X} = X | \neg\Omega_{\leq j}, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j;$$

and

$$\tilde{Y} = Y | \neg\Omega_{\leq j}, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j.$$

Then Alice wants to send Bob the message $M(x) = M_{j+1}(x, \tilde{m}_1, \ldots, \tilde{m}_j)$, which she does once again by running the protocol σ of Theorem 13 with the same error parameter as before; of course, Alice and Bob have no way of knowing whether an error happened or not, but they always assume that it did not, and the protocol σ is played as if the inputs were given according to $\tilde{X}$ and $\tilde{Y}$.

Now we see that, conditioned on $\neg\Omega_{\leq j}$ and on σ producing the correct message $M(x)$ (which together form the event $\neg\Omega_{\leq j+1}$), then σ uses, on average, a constant number of rounds, and at most $O(H(M(\tilde{X})|\tilde{Y}) + \log \varepsilon^{-1})$ bits of communication. But, since conditioning on some event can only decrease the entropy of a given distribution, we find that

$$\begin{aligned} H(M(\tilde{X})|\tilde{Y}) &= H(M_{j+1}(\tilde{X}, m_1, \ldots, m_j)|\tilde{Y}) \\ &= H(M_{j+1}(X, m_1, \ldots, m_j)|Y, \neg\Omega_{\leq j}, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j) \\ &\leq H(M_{j+1}(X, m_1, \ldots, m_j)|Y, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j) \\ &= I(X : M_{j+1}|Y, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j). \end{aligned}$$

Our event Ω will of course be $\Omega_{\leq q} = \Omega_1 \cup \ldots \cup \Omega_q$. From the previous equation we find that, conditioned on $\neg\Omega$, the average communication of $\tilde{\rho}$ for round $j+1$ is at most

$$\mathbb{E}_{\tilde{m}_1,\ldots,\tilde{m}_j}[O(I(X : M_{j+1}|Y, \Pi_{\leq j} = \tilde{m}_1, \ldots, \tilde{m}_j) + \log \varepsilon^{-1}] = I(X : M_{j+1}|Y, \Pi_{\leq j}) + \log(\delta^{-1}q),$$

which summed for every round[6] will total $O(\mathsf{IC}_\mu(\pi) + q \log(\delta^{-1}q))$. Finally, from a union bound it follows that the error probability $\Pr[\Omega_{\leq j}] \leq j\varepsilon$, and hence the total error probability will be at most $q\varepsilon = \delta$. □

As a corollary we get Theorem 4.

[6] And taking into account that the situation is entirely symmetric when it is Bob's turn to communicate.

Proof of Theorem 4. Take the protocol $\tilde{\rho}$ given by Theorem 14, say with error $\delta/3$, and let ρ be the same protocol *truncated* to $O(\delta^{-1}(I + q\log(\delta^{-1}q)))$ bits of communication and $O(\delta^{-1}q)$ rounds; i.e., ρ proceeds as in $\tilde{\rho}$, but aborts if the number of used rounds or communication is about to exceed said bounds. By Markov's inequality it holds that the probability that $\tilde{\rho}$ ever uses more rounds or more communication is at most $\delta/3$, hence by a union bound ρ will successfully simulate π except with error probability δ. □

A.2 A variant of Theorem 4 that uses even less rounds

A.3 Proof of Theorem 3

We will use the following protocol in the proof of Theorem 3:

Lemma 15. *(REF) There is a randomized public-coin protocol with communication complexity $O(\log(k/\varepsilon))$ such that on input two k-bit strings x, y, it outputs the first index $i \in [k]$ such that $x_i \neq y_i$ with probability at least $1 - \varepsilon$, if such an i exists.*

Proof of Theorem 3. The protocol ρ works as follows. On inputs x and y, Alice and Bob first pick the shared randomness r in the same way as in protocol π. Then Alice forms the set

$$\Pi^{(a)} = \{t \in \{0,1\}^{\mathsf{CC}(\pi)} | \Pr[\Pi = t|x, r] \geq 2^{-25\delta^{-2}I}\},$$

and Bob forms the symmetric set

$$\Pi^{(b)} = \{t \in \{0,1\}^{\mathsf{CC}(\pi)} | \Pr[\Pi = t|y, r] \geq 2^{-25\delta^{-2}I}\}.$$

By simple counting, it is clear that each set has no more than $2^{25\delta^{-2}I}$ transcripts. Furthermore, since $\mathsf{IC}_\mu(\pi) = H(\Pi|X, R) + H(\Pi|Y, R) \leq I$, then by the Markov inequality it will hold — with probability $\geq 1 - \frac{\delta}{5}$ over the choice of x and r — that

$$H(\Pi|x, r) \leq \frac{5}{\delta} I.$$

Conditioned on this event and using Fact 2, then we will have $\pi(x, y, r) \in \Pi^{(a)}$ — with probability $\geq 1 - \frac{\delta}{5}$ over the choice of y. In this way it can be seen that the event $\pi(x, y, r) \in \Pi^{(a)} \cap \Pi^{(b)}$, which we denote by Ω, will hold with probability $\geq 1 - \frac{4}{5}\delta$ over the choice of x, y and r. Indeed, conditioned on Ω, it will hold that $\Pi^{(a)} \cap \Pi^{(b)} = \{\pi(x, y, r)\}$, because every string in $\Pi^{(a)}$ is of the form $\pi(x, y', r)$ for some y', and every string in $\Pi^{(b)}$ is of the form $\pi(x', y, r)$ for some x', and clearly if $t = \pi(x, y', r) = \pi(x', y, r)$, then it must hold that $t = \pi(x, y, r)$. Furthermore, if Ω fails to hold, then $\Pi^{(a)} \cap \Pi^{(b)} = \varnothing$.

This suggests that we might find $\pi(x, y, r)$, by running a generic protocol to uncover the intersection of $\Pi^{(a)}$ and $\Pi^{(b)}$, but it is know that such a protocol will require $\Omega(|\Pi^{(a)}| + |\Pi^{(b)}|)$ communication (REF Wigderson and the other guy). So we must necessarily exploit the structure of the Π's by looking inside

the protocol. Let $\ell = \log|\Pi^{(a)}| + \log|\Pi^{(b)}|$, and recall that $C = \mathsf{CC}(\pi)$. We will now conclude the proof by showing that Alice and Bob can determine the intersection of $\Pi^{(a)} \cap \Pi^{(b)}$ by communicating $O(\delta^{-1}\ell \log(\ell C)) = O(\delta^{-3} I \log C)$ many bits, and with a chance of failure bounded by $\delta/5$.

To do this, Alice organizes $\Pi^{(a)}$ into the prefix-tree $T^{(a)}$, as follows: the root is the largest common prefix (lcp) of the transcripts in $\Pi^{(a)}$, and the remaining nodes are defined inductively. If we have node τ, then we let its children be $\tau 0 \tau_0$ and $\tau 1 \tau_1$, where τ_i is the lcp of the transcripts in $\Pi^{(a)}$ beginning with τi.

The tree $T^{(a)}$ is a possibly quite unbalanced binary tree, whose leaves are the transcripts $t \in \Pi^{(a)}$. Of course, Bob has a similar tree $T^{(b)}$, and the protocol ρ is required to find a common leaf with little communication. Note that $\pi(x, y, r)$ is in both trees, and hence the root of both $T^{(a)}$ and $T^{(b)}$ is a prefix of $\pi(x, y, r)$. In the descriptions below, we will use $t^{(a)}$ (and $t^{(b)}$) to designate a leaf of $T^{(a)}$ (resp. $T^{(b)}$), and $\tau^{(a)}$ ($\tau^{(b)}$) to designate an arbitrary node of $T^{(a)}$ (resp. ($T^{(b)}$).

Alice and Bob will proceed in stages to find a common leaf of $T^{(a)}$ and $T^{(b)}$:

(1) At the beginning of stage s, they will have agreed that certain nodes in their respective trees — $\tau^{(a)}(s) \in T^{(a)}$ and $\tau^{(b)}(s) \in T^{(b)}$ — are prefixes of $\pi(x, y, r)$. In the first stage, $\tau^{(a)}(1)$ and $\tau^{(b)}(1)$ will be the roots of the trees.

(2) Then Alice picks a *candidate leaf* $t^{(a)}$ that is a successor of $\tau^{(a)}(s)$, and Bob picks a candidate leaf $t^{(b)}$ that is a successor of $\tau^{(b)}(s)$.

(3) They use the protocol of Lemma 15 to find the least position j for which $t^{(a)}_j \neq t^{(b)}_j$.[7] If there is no such j then they both have $t^{(a)} = t^{(b)} = \pi(x, y, r)$, and the simulation terminates.

(4) Now, if in the protocol π it was Alice's turn to communicate bit j, then — in protocol ρ — Alice sets $\tau^{(a)}(s+1) = \tau^{(a)}(s)$, and Bob will set $\tau^{(b)}(s+1)$ as explained in the next item. Then they proceed to the next stage.

(5) Bob looks at the node $\hat{\tau}^{(b)}$ given the first $j-1$ bits of his candidate leaf $t^{(b)}$. Now suppose that $t^{(b)}$ is a left successor of $\hat{\tau}^{(b)}$. Then Bob will set $\tau^{(b)}(s+1)$ to be the right child of $\hat{\tau}^{(b)}$. The node $\tau^{(b)}(r+1)$ is called the "flip" of $t^{(b)}$ at position j.

Note that $\hat{\tau}^{(b)}$ must be a node in the tree $T^{(b)}$, since $\pi(x, y, r)$ is prefixed by $\hat{\tau}^{(b)}$ and has a different bit than $t^{(b)}$ on the j-th position (namely, the j-th bit of $t^{(a)}$). Because every left child of $\hat{\tau}^{(b)}$ will have the same bit as $t^{(b)}$ at position j, then $\pi(x, y, r)$ can not be a left successor of $\hat{\tau}^{(b)}$. Hence $\pi(x, y, r)$ must be a successor of $\tau^{(b)}(s+1)$, and the property required by item (1) is preserved for stage $s+1$. Of course, a symmetric argument holds when $t^{(b)}$ is a right successor of $\hat{\tau}^{(b)}$, and a similar reasoning is applied if in protocol π it was Bob's turn to communicate bit j (then it will be Alice who changes her prefix).

[7] They will make sure that the protocol of Lemma 15 succeeds with a certain error parameter ε we will give later.

The property required by item (1) ensures that, conditioned on Ω, Alice and Bob will eventually agree on π itself. All that is left for us to specify is how Alice and Bob pick their candidate. We will show that for any binary prefix tree T, and for any prefix $\tau \in T$, there is a way of picking a candidate leaf t extending τ such that any "flip" of t will have at most half as many successors as τ. Then this implies that the number of successors of either $\tau^{(a)}(s)$ or $\tau^{(b)}(s)$ will halve at each stage, and hence our simulation ends in at most ℓ stages. If we set the error parameter of Lemma 15 to be $\frac{\delta}{5\ell}$, then each stage uses $O(\delta^{-1}\log(\ell C))$ many bits of communication,[8] and we will successfully find $\pi(x, y, r)$ with probability $\frac{\delta}{5}$, as we intended.

The candidate is picked as follows: we let $\tau_1 = \tau$ to begin with; if we have already defined τ_i, then we choose τ_{i+1} to be whichever child (left or right) has more successors in T, (arbitrarily) opting for the left child if both numbers are the same. When we reach a leaf, this is our candidate. Define S_i to be the number of successors of τ_i. Suppose τ' is a flip of our candidate, i.e., τ' is the one child of τ_i, for some i for which τ_{i+1} is the other child. Then by our choice of τ_{i+1}, τ' has at most $S_i/2 \leq S/2$ many successors, as intended. □

A.4 Proof of Theorem 9

Proof. Suppose that we could compute $f^{\otimes k}$ in $o(q)$ rounds with kC bits of communication with error ε. Then by Theorem 8, there is a protocol π for computing f using kC bits of communication and C bits of information. Hence by Corollary 6, there is a public-coin protocol ρ for computing f using $O(C + q\log q(n + C))$ many bits of communication, in $O(q)$ rounds, with error δ (the precise constant depends on δ).

The particular case follows by applying this result with

$$C = \frac{1}{d}\left(R^{dq}_{\delta}(f) - O\left(q\log\left(qn + qR^{dq}_{\delta}(f)\right)\right)\right).$$

□

B Proofs for Section 4

Proof of Lemma 11. We show the existence of such a graph using a probablistic argument. Let $A = \{u_1, \ldots, u_{kN}\}$ and $B = \{v_1, \ldots, v_n\}$. Construct a random graph G by choosing d random neighbors independently for each $u \in A$. For any $A' \subseteq A$ of size N, let $E_{A'}$ be the event that $G_{A' \cup B}$ does *not* have a matching of size $N(1-\delta)$, and let $B_{AD} := \bigvee_{A'} E_{A'}$. Note that the lemma holds if $\Pr[B_{AD}] < 1$.

Next, we bound $\Pr[E_{A'}]$. Consider the following procedure for generating a matching for $G_{A' \cup B}$:

[8] Actually the dependence on δ^{-1} is also an additive logarithmic term, but we upper bound it like this aiming at simplicity of the final expression.

FIND-MATCHING

```
1  M ← ∅
2  V ← ∅
3  for i ← 1 to N
4       if N(u_i) ⊈ V
5            pick arbitrary v_i ∈ N(u_i) \ V
6            M ← M ∪ {(u_i, v_i)}
7            V ← V ∪ {v_i}
8  return M
```

Let $X_1, \ldots, X_N$ be indicator variables for the event that the matching increased at step i, and let $Y_1, \ldots, Y_N$ to be i.i.d. random coins with $\Pr[Y_i = 1] = e^{-d\delta}$. Define $BAD_{A'}$ to be the event that $\sum_i X_i < N(1-\delta)$. In other words, $BAD_{A'}$ is the event that FIND-MATCHING fails to return a large enough matching for $G_{A' \cup B}$. For any i, the matching fails to increase at step i only when all neighbors of u_i have already been matched. It follows that

$$\Pr[X_i = 0] = \left(\frac{\sum_{j<i} X_j}{N} \right)^d .$$

Furthermore, assuming that a large matching has not been found by step i, we have

$$\Pr[X_i = 0] < (1-\delta)^d < \Pr[Y_i = 1] . \tag{1}$$

In fact, we claim the following.

Claim 1. $\Pr[E_{A'}] \leq \Pr[BAD_{A'}] \leq \Pr[\sum_i Y_i > \delta N]$.

It remains to bound this latter probability. We use the following claim, with $p := e^{-d\delta}$.

Claim 2. *Let $Y_1, \ldots, Y_N$ be i.i.d. biased coins, with $\Pr[Y_i = 1] = p < \delta < 1$. Then,*

$$\Pr\left[\sum Y_i > \delta N \right] < \exp(\delta N (1 + \ln(p/\delta))) .$$

Next, we bound the number of subsets $A' \subset A$ of size N, with the following claim.

Claim 3. *There are at most $\exp(N(1+\ln k))$ subsets of A of size N.*

Taking the two claims together, we have

$$\begin{aligned} \Pr[BAD] &\leq \exp\left(N(1+\ln k)\right) \cdot \exp\left(\delta N(1+\ln(p/\delta))\right) \\ &= \exp\left(N + N\ln k + \delta N + \delta N \ln(1/\delta) + \delta N \ln p\right) \\ &= \exp\left(N + N\ln k + \delta N + \delta N \ln(1/\delta) - d\delta^2 N\right) \\ &< 1 , \end{aligned}$$

where the final inequality uses $d = (2 + \ln k)/\delta^2 + \ln(1/\delta)/\delta$. □

Now let us prove the claims.

Proof of Claim 1. For a string $x \in \{0,1\}^N$, let $x_{\leq i}$ denote the substring $x_1 \cdots x_i$, and call x bad if $|x| < N(1-\delta)$. For $0 \leq j \leq n$, consider the random variable

$$D^{(j)} = X_1 \ldots X_j (1 - Y_{j+1}) \ldots (1 - Y_N).$$

Now notice that for any string v of length i, it holds that $\Pr[D_i = v] = \Pr[D_{i+1} = v]$. We have two cases:

- If $|v| \geq N(1-\delta)$, then

$$\Pr[D^{(i)} \text{ is bad} | D^{(i)}_{\leq i} = v] = \Pr[D^{(i+1)} \text{ is bad} | D^{(i+1)}_{\leq i} = v] = 0;$$

- If $|v| < N(1-\delta)$, then from equation (1) we get

$$\begin{aligned} \Pr[D^{(i)}_{i+1} = 1 | D^{(i)}_{\leq i} = v] &= \Pr[Y_{i+1} = 0 | \vec{X}_{\leq i} = v] \\ &> \Pr[X_{i+1} = 1 | \vec{X}_{\leq i} = v] \\ &= \Pr[D^{(i+1)}_{i+1} = 1 | D^{(i+1)}_{\leq i} = v] = 0. \end{aligned}$$

So in either case we conclude that

$$\Pr[D^{(i+1)} \text{ is bad}] \leq \Pr[D^{(i)} \text{ is bad}],$$

and the claim follows. □

Proof of Claim 2. Let $Y := \sum Y_i$, and let $\mu := \mathbb{E}[Y]$. Note that $\mu = pN$. Also, let $\psi := \delta/p - 1$. Using the multiplicative version of the Chernoff bound, we have

$$\begin{aligned} \Pr[\sum Y_i > \delta N] &= \Pr[Y > pN \cdot (\delta/p)] \\ &= \Pr[Y > \mu(1+\psi)] \\ &< \left(\frac{e^{\psi}}{(1+\psi)^{(1+\psi)}} \right)^{\mu} \\ &= \exp\left(\mu \left(\frac{\delta}{p} - 1 - \frac{\delta}{p} \ln(\frac{\delta}{p}) \right) \right) \\ &= \exp\left(pN \frac{\delta}{p} \left(1 - \frac{p}{\delta} - \ln \delta + \ln p \right) \right) \\ &= \exp\left(\delta N - pN + \delta N \ln(1/\delta) + \delta N \ln p \right) \\ &< \exp\left(\delta N \left(1 + \ln(1/\delta) + \ln p \right) \right). \end{aligned}$$

□

Proof of Claim 3. There are $\binom{kN}{N}$ subsets of A of size N. By Stirling's Formula, we have

$$\binom{kN}{N} \leq \frac{(kN)^N}{N!} \leq \left(\frac{kNe}{N} \right)^N = \exp(N(1 + \ln k)).$$

□

Proof of Theorem 10. Let $k = \log|\mathcal{X}|$ and $N = |\mathcal{R}|$. Assume without loss of generality that $\mathcal{M} = M(\mathcal{X}, \mathcal{R})$; then $|\mathcal{M}| \leq 2^k N$. Now let G be $(2^k N, N, d, \delta)$-matching graph having $\mathcal{M}$ as its left set and $\mathcal{R}$ as its right set, for $\delta = \frac{\varepsilon}{2k^2}$. For these parameters, we are assured by Lemma 11 that such a matching graph exists having left-degree $d \leq O(\frac{k^5}{\varepsilon})$.

We construct the new protocol M' as follows:

For each $x \in \mathcal{X}$ let $\mathcal{M}_x = M(x, \mathcal{R})$ be the set of messages that might be sent on input x. Noticing that $|M_x| = N$, consider a partial G-matching between $\mathcal{M}_x$ and $\mathcal{R}$ pairing all but a δ fraction $\mathcal{M}_x$; then define a bijection $\phi_x : \mathcal{R} \to \mathcal{M}_x$ by setting $\phi_x(r) = m$ if $\{m, r\}$ is an edge in the matching, and pairing the unmatched m and r's arbitrarily (possibly using edges not in G). Finally, set $M'(x, r) = \phi_x(r)$.

Since $M'(x, r) = M(x, \sigma(r))$ for some permutation σ, then it is clear that M and M' generate the same transcript.

Now we prove that M' does not leak much more information than M. Since $I(X : M', R) - I(X : M) = H(X|M) - H(X|M', R)$, it suffices to show that $H(X|M', R) \geq H(X|M) - O(\log k)$.

A tripple $(x, r, M(x, r))$ will be called a *cell* of message $m = M(x, r)$. A given pair $\{m, r\}$ will be called *good* when $\{m, r\}$ is an edge of G, and a cell (x, r, m) is called *good* if $\{m, r\}$ is good. Also, let us call a message m *good* if its good cells make up at least a $1 - \frac{1}{k}$ fraction of the probability mass. We will say "bad" as a shorthand for "not good." The following claim will be proven later:

Claim 4. *For our choice of parameters,* $\Pr[M'(X, R)$ *is bad*$] < \frac{1}{k}$.

Now, if R_m denotes the distribution R conditioned on $M'(X, R) = m$, then

$$H(X|M', R) = \mathbb{E}_{m \sim M'(X,R)}[H(X|M' = m, R_m)]. \tag{2}$$

For each fixed m, the right-hand entropy equals

$$H(X|M' = m, R_m) = H(X|M' = m) - I(X : R|M' = m). \tag{3}$$

But $I(X : R|M' = m) = H(R_m) - H(R_m|M' = m, X) = H(R_m)$, because since M' is 1–1, if we know m and x then r is completely determined. Now because M and M' are equidistributed for every x, then we get:

$$H(X|M', R) = H(X|M) - H(R|M').$$

All that is left to do is bound $H(R|M')$. For any fixed m we have $H(R_m) \leq k$, because r is a function of m (which is given) and x (which is k-bits long). Hence,

$$H(R|M') \leq \Pr[M' \text{ is good}]\, \mathbb{E}_{\text{good } m}[H(R_m)] + \Pr[M' \text{ is bad}]k.$$

And for good m,

$$H(R_m) \leq \Pr[\{m, R_m\} \text{ is good}] H(R_m|\text{good } \{m, R_m\}] + \Pr[\{m, R_m\} \text{ is bad}]k.$$

We now have that $\Pr[M' \text{ is bad}]$ and $\Pr[\{m, R_m\} \text{ is bad}]$ are both less than $\frac{1}{k}$, by Claim 4 and assuming that m is good, respectively. Furthermore, conditioned

on $\{m, R_m\}$ being good, the support of R_m is at most $d = O(k^5)$, and hence $H(R_m|\text{good } \{m, R_m\})$ is at most $\log d = O(\log k)$; hence we obtain

$$H(X|M', R) \geq H(X|M) - O(\log k).$$

□

Proof. (Claim 4) Suppose that $\Pr[M'(X, R) \text{ is bad}] > \frac{1}{k}$. Then the probability that $(X, R, M'(X, R))$ is a bad cell is at least

$$\Pr[M'(X, R) \text{ is bad}] \Pr[(X, R, M'(X, R))\text{is bad}|M'(X, R) \text{ is bad}] > \frac{1}{k^2}.$$

But then there must exist a choice of x such that $\Pr[(x, R, M'(x, R)) \text{ is bad}] > \frac{1}{k^2}$, which implies that, for this x, there is a $\frac{1}{k^2}$ fraction of the $(r, M'(x, r))$ pairs that are not edges of G, and hence not part of the matching. But this contradicts the fact that our matching gives at most a $\delta < \frac{1}{k^2}$ fraction of unmatched vertices.

□

C Proofs of Section 5

Proof of Theorem 12. We think of $M(\cdot, \cdot)$ as a table, which we will call *the M-table*, where the inputs $x \in \mathcal{X}$ are the rows and the random choices $r \in \mathcal{R}$ are the columns, and fix some ordering $r_1 < r_2 < \ldots$ of $\mathcal{R}$. The second part $J(x, r)$ of $\tilde{M}$ will be set to the number of times $M(x, r)$ appeared in the same row up to the column r, i.e.,

$$J(x, r) = |\{r' \leq r | M(x, r') = M(x, r)\}|.$$

From this point onwards, let us fix the message m, and denote the conditional distribution $X|M(X, R) = m$ with X_m, $R|M(X, R) = m$ with R_m, and the distribution $J(X, R)|M(X, R) = m$ with J_m. The supports of X_m and R_m will be denoted $\mathcal{X}_m$ and $\mathcal{R}_m$, respectively. We will settle the theorem by proving that the following holds, regardless of our choice of m:

$$I(X_m : J_m) = H(X_m) - H(X_m|J_m) = O(\log \log |\mathcal{X}||\mathcal{R}|).$$

To prove this we will reduce it to a question about certain combinatorial objects that we non-descriptionally call *shapes*. Let us start by making the simplifying assumption that the distribution X over Alice's inputs is uniform over $\mathcal{X}$. What do X_m and J_m look like in this particular case? Well if we let $w_x = |\{r \in \mathcal{R} | M(x, r) = m\}|$ be the number of m-entries in row x, and $T = \sum_x w_x$ be the total number of m-entries in the M-table, then

$$\Pr[X_m = x] = \frac{w_x}{T}.$$

Also, if we let $h_i = |\{x \in \mathcal{X}_m | w_x \geq i\}|$ then it holds that $T = \sum_i h_i$ and

$$\Pr[J_m = i] = \frac{h_i}{T}.$$

So we can represent both distributions X_m and J_m by a combinatorial object, called a *simple shape*, in the following way: our simple shape S_m will have $|\mathcal{X}_m|$ rows of *squares*, precisely w_x squares on row x, for a total of T squares, and the i-th column will then have h_i squares, like so:

1	2	3	...	w_1		
1	2	3	4	5	...	w_2
...						
1	2	...	w_N			

The simple shape S_m has the property that if we pick a square at random the row will be distributed like X_m and the column will be distributed like J_m. In general, an arbitrary simple shape S is a set of (x, i) squares obeying certain rules (cf. Section C.1), we may define w_x, h_i, and T as before. Then the *entropy loss* of S is

$$el(S) = \sum_x \frac{w_x}{T} \log \frac{T}{w_x} - \sum_i \frac{h_i}{T} \log h_i,$$

where x ranges over the rows and i over the columns. We now have that $I(X_m : J_m) = el(S_m)$. All we need to do is prove good bounds on $el(S)$.

To deal with the general case, where X can be arbitrarily distributed, we will need a more general notion of shape. The new shapes similar, but we further divide each square into smaller pieces, and think of each piece as a *unit of probability mass*. A shape is now a pair $\mathcal{S} = (S, s)$ of a simple shape S together with a function $s : S \to \mathbb{N}$ such that $s(x, i)$ is the number of pieces in the (x, i)-square of S. We define $w_x = \sum_i s(x, i)$, $h_i = \sum_x s(x, i)$ and $T = \sum_{x,i} s(x, i)$. This induces a distribution $(\tilde{X}, \tilde{J})$ over the squares of S, obtained by picking a piece uniformly at random from $\mathcal{S}$, i.e. $\Pr[(\tilde{X}, \tilde{J}) = (x, i)] = \frac{s(x,i)}{T}$; the entropy loss is now

$$el(\mathcal{S}) = H(\tilde{X}) - H(\tilde{X}|\tilde{J}).$$

We will be able to show (Theorem 17) that $el(\mathcal{S}) \leq O(\log \log T)$. Therefore, to finish the proof of the theorem, it suffices to show that there is some shape $\mathcal{S}_m$ such that $I(X_m : J_m) = el(\mathcal{S}_m) + O(1)$, and for which $T = O((|\mathcal{X}||\mathcal{R}|)^{O(1)})$.

Let S denote the support of (X_m, J_m). Then the squares of $\mathcal{S}_m$ are the pairs in S: row x of $\mathcal{S}_m$ has v_x squares $(x, 1), \ldots, (x, v_x)$, where $v_x = |\{r \in \mathcal{R} | M(x, r) = m\}$ is the number of random choices which can result in sending message m on input x. To specify the number of units of mass in each of these squares, we first construct a suitable discretization of the distribution (X_m, J_m). Noticing that $\Pr[(X_m, J_m) = (x, j)]$ is the same for each j in the support of J_m, define $\tilde{p}_x$ as the integer such that

$$P_{x,j} = \Pr[(X_m, J_m) = (x, j)] = \frac{\tilde{p}_x + d_x}{V} \qquad d \in [0, 1),$$

for a given value V to be given later. Then square (x, j) will have exactly $s(x, j) = \tilde{p}_x$ units of probability mass. This induces a pair of distributions

$(\tilde{X}_m, \tilde{J}_m)$ on S, with

$$\tilde{P}_{x,j} = \Pr[(\tilde{X}_m, \tilde{J}_m) = (x,j)] \quad = \quad \frac{\tilde{p}_x}{\sum_{(x,j)\in S} \tilde{p}_x}. \tag{4}$$

Since $\tilde{p}_x = P_{x,j}V - d_x$, we get the bounds:

$$P_{x,j} - \frac{1}{V} \leq \frac{P_{x,j}V - d_x}{V} \leq \tilde{P}_{x,j} = \frac{P_{x,j}V - d_x}{V - \sum_{(x,j)\in S} d_x} \leq \frac{P_{x,j}V}{V - |S|}$$

To bound the right-hand side, we take the Taylor series expansion of $\frac{az}{z-b}$ at $z = +\infty$, which is

$$a + \sum_{i=1}^{\infty} \frac{ab^i}{z^i}.$$

From this we find, for $a = P_{x,j}$, $b = |S|$, $V \geq 4|S||\mathcal{X}|^2|\mathcal{R}|^2$, that:

$$|P_{x,j} - \tilde{P}_{x,j}| \leq \frac{1}{2|\mathcal{X}|^2|\mathcal{R}|^2}.$$

Hence the following:

Claim 5. *Let $\varepsilon = \frac{1}{|\mathcal{X}||\mathcal{R}|}$; then X_m and $\tilde{X}_m$ are ε-close; J_m and $\tilde{J}_m$ are ε-close; and, for every j, $X_m|J_m = j$ and $\tilde{X}_m|\tilde{J}_n = j$ are ε-close (in statistical distance).*

Now we make use of the following fact:

Fact 16. *For any two distributions A, B over the same universe $\mathcal{U}$, it holds that*

$$|H(A) - H(B)| \leq \log(|\mathcal{U}|)\delta(A,B) + 1$$

Where $\delta(A,B)$ is the statistical difference between A and B.

Together with the former claim, this implies that $H(X_m) = H(\tilde{X}_m) + O(1)$, and that, for every j, $H(X_m|J_m = j) = H(\tilde{X}_m|\tilde{J}_m = j) + O(1)$, and hence that $H(X_m|J_m) = H(\tilde{X}_m|\tilde{J}_m) + O(1)$. Therefore,

$$I(X_m : J_m) = I(\tilde{X}_m : \tilde{J}_m) + O(1) = el(\mathcal{S}_m) + O(1),$$

and this concludes the proof. □

C.1 Shapes

Definition 6. *A* simple shape *is a finite set S of pairs of positive natural numbers, called* squares*, obeying the following:*

1. *(Row axiom) If $(x,i) \in S$ and $i \geq 2$ then $(x, i-1) \in S$.*

A shape *$\mathcal{S}$ is a pair (S, s) of a simple shape together with a function $s : \mathbb{N}^2 \to \mathbb{N}$, obeying the following:*

1. *(Mass is in squares)* $s(x,i) = 0$ *if* $(x,i) \notin S$ *and* $s(x,i) > 0$ *if* $(x,i) \in S$.

If $s(x,i) = k$*, then we say that* $\mathcal{S}$ *has* k units of mass *in the* (x,i)*-square. Associated with* $\mathcal{S}$ *we define the quantities:* $T = \sum_{(x,i)\in S} s(x,i)$ *is the* total mass *in* $\mathcal{S}$*; for each* x*,* $w_x = \sum_{i:(x,i)\in S} s(x,i)$ *is the* row-mass*; for each* i*,* $h_i = \sum_{x:(x,i)\in S} s(x,i)$ *is the* column-mass.

A shape $\mathcal{S} = (S, s)$ *is called* regular *if it obeys the following:*

- *(Regularity axiom)* $s(x, \cdot)$ *is constant within each row.*

We can visualize a simple shape as a a finite collection of squares, arranged in left-justified rows. A shape is simply a filling of the squares with positive natural numbers, and the shape is regular when in each row we have the same positive number in each square. An example of a regular shape is:

10	10	10	10	10
5	5	5		
15	15	15	15	
40				
50	50			

We now define some further notions associated with a given shape; for simple shapes, the same notions are valid by setting $s(x,i) = 1$.

Definition 7. *A given shape* $\mathcal{S} = (S, s)$*, induces an* associated distribution (X, I) *on the squares of* S*, given by:*

$$\Pr[(X, I) = (x, i)] = \frac{s(x,i)}{T}.$$

The entropy loss *of* $\mathcal{S}$ *is now defined to be* $el(\mathcal{S}) = I(X : I) = H(X) - H(X|I)$.

The precise equation for $el(\mathcal{S})$ is:

$$el(\mathcal{S}) = \sum_x \frac{w_x}{T} \log \frac{T}{w_x} - \sum_i \frac{h_i}{T} \left(\sum_x \frac{s(x,i)}{h_i} \log \frac{h_i}{s(x,i)} \right),$$

and this can be expanded to:

$$el(\mathcal{S}) = \log T - \frac{1}{T} \left(\sum_x w_x \log w_x + \sum_i \left(h_i \log h_i - \sum_x s(x,i) \log s(x,i) \right) \right).$$

Our main theorem in this section is the following:

Theorem 17. *For any regular shape* $\mathcal{S}$*,* $el(\mathcal{S}) \leq O(\log \log T)$.

The proof is rather intricate and will be divided into several parts. We will first show that $\mathcal{S}$ can be approximated by a fairly rough shape $\mathcal{S}'$, so that $\mathcal{S}'$ is "close" to $\mathcal{S}$, and this will in particular imply that $el(\mathcal{S})$ and $el(\mathcal{S}')$ are about the same. Then $el(\mathcal{S}')$ itself will be shown to be small. The notion of "close" is based on the following definitions:

Definition 8 (distance)**.** *Let $\mathcal{G}$ be an infinite weighted graph whose vertices are the shapes. An edge $\{\mathcal{S}, \mathcal{S}'\}$ is present if $\mathcal{S}'$ can be obtained from $\mathcal{S}$ by adding or removing one unit of mass, i.e., $s'(y,j) = s(y,j) - 1$ (or $+1$) for some square (y,j). The weight of the edge $\{\mathcal{S}, \mathcal{S}'\}$ is $|\, el(\mathcal{S}) - el(\mathcal{S}')|$.*

The distance between two shapes $\delta(\mathcal{S}, \mathcal{S}')$ *is now the weight of the shortest (least-weight) path between S and S'.*

Note that when $\mathcal{S}, \mathcal{S}'$ are not neighbors, $\delta(\mathcal{S}, \mathcal{S}')$ does not necessarily equal $|\, el(\mathcal{S}) - el(\mathcal{S}')|$; by the triangle inequality, however, it does hold that $|\, el(\mathcal{S}') - el(\mathcal{S})| \leq \delta(\mathcal{S}, \mathcal{S}')$.

Suppose that $\{\mathcal{S}, \mathcal{S}'\}$ is an edge of $\mathcal{G}$, say $s'(x,i) = s(x,i) + 1$. Then:

$$\begin{aligned} el(\mathcal{S}) - el(\mathcal{S}') &= \log T - \log(T+1) - \frac{1}{T(T+1)}\,(W + V) \\ W &= \sum_x (T+1) w_x \log w_x - T w'_x \log w'_x \\ V &= \sum_i (T+1) h_i H(X|i) - T h'_i H(X'|i) \end{aligned}$$

Let us develop the terms W, taking into account the cancelations arising from the fact that $w'_x = w_x$ for all $x \neq z$, and $w'_z = w_z + 1$.

$$W = \sum_x w_x \log w_x + T w_z (\log w_z - \log(w_z + 1)) - T \log(w_z + 1) = O(T \log T)$$

Now we proceed similarly for V, using the fact that $h'_i = h_i$ for all $i \neq j$, and $h'_j = h_j + 1$ (this implies that $X'|i = X|i$ for $i \neq j$).

$$V = \sum_i h_i H(X|i) + T h_j (H(X|j) - H(X'|j)) - T H(X'|j)$$

The partial term $H(X|j) - H(X'|j)$ can be expanded in a similar way to show an upper bound of $O(\frac{\log h_j}{h_j})$, and thus $V = O(M \log M)$ also. Recalling that $\ln T - \ln(T+1) < \frac{2}{T}$ for positive T, it can be seen that:

$$|\, el(\mathcal{S}) - el(\mathcal{S}')| = O\left(\frac{\log T}{T}\right)$$

From this it follows:

Corollary 18. *If there is a path of length ℓ between $\mathcal{S}$ and $\mathcal{S}'$ in $\mathcal{G}$, and $T \leq T'$, then $\delta(\mathcal{S}, \mathcal{S}') \leq O(\frac{\ell \log T'}{T})$.*

Now we are ready to prove Theorem 17:

Proof. (Theorem 17) Let $\mathcal{S} = (S, s)$ be a given regular shape with total mass T. We will construct a larger regular shape $\mathcal{S}' = (S', s)$ with the following properties:

(i) The rows of $\mathcal{S}'$ can be partitioned into $\ell = O((\log T)^4)$-many *blocks* (sets of rows) $\mathcal{X}_1, \ldots, \mathcal{X}_\ell$, such that the rows x of each block have the same number of squares.

(ii) There is a path between $\mathcal{S}'$ and $\mathcal{S}$ of length $\frac{T}{\log T'}$.

From item (i) it follows that $el(\mathcal{S}')$ is at most $O(\log\log T)$. To see this, consider the distributions (X', I') associated with $\mathcal{S}'$. Then X' is of the form (K, X_K), where K first chooses a block $\mathcal{X}_k$ (with probability $\frac{1}{T}\sum_{x\in\mathcal{X}_k} w_x$), and then X_k chooses a row inside $\mathcal{X}_k$. The entropy loss is:

$$el(\mathcal{S}') = I(X' : I') = I(K : I') + I(X_K : I'|K).$$

By assumption, let v_k be the number of squares $(x, i) \in S'$, for the rows $x \in \mathcal{X}_k$. Then because $\mathcal{S}'$ is regular, we have that

$$\Pr[X_k = x] = \frac{v_k s'(x, 1)}{T} = \frac{v_k s'(x, i)}{T}$$

for every k and i. But then $X_k|I = i$ is exactly the same distribution as X_k, and thus $I(X_K : I'|K) = 0$. Therefore we conclude that $el(\mathcal{S}') = I(K : I') \leq H(K) \leq \log \ell \leq O(\log\log T)$.

Now from item (ii) and Corollary 18, it holds that $el(\mathcal{S}) \leq el(\mathcal{S}') + O(1) = O(\log\log T)$.

Here is how we construct $\mathcal{S}'$. From the regularity of $\mathcal{S}'$, let $d_x = s(x, 1)$ denote the number of units of mass in each square of row x, and let $v_w = \frac{w_x}{d_x}$ denote the number of squares in row x. Let $B = \left(1 + \frac{1}{6(1+\log T)}\right)$. Now for every $1 \leq k_1 \leq k_2 \leq 6(\log T)^2$ let $\mathcal{X}_{k_1,k_2}$ be the subset of rows x of $\mathcal{S}$, such that:

$$v_x > 2\log T \quad \text{and} \quad B^{k_1-1} \leq d_x < B^{k_1} \quad \text{and} \quad B^{k_2-1} \leq w_x < B^{k_2}.$$

Also, for $1 \leq k \leq 3(1 + \log T)$, let $\mathcal{X}_k$ denote the rows of $\mathcal{S}$ that have exactly $v_x = k$ squares. It is clear that every row of $\mathcal{S}$ must be in one of the blocks $\mathcal{X}_k$ or $\mathcal{X}_{k_1,k_2}$. The shape $\mathcal{S}'$ assigns $\lceil B^{k_2-k_1-1}\rceil$ squares to the rows in $\mathcal{X}_{k_1,k_2}$, and $s(x, i) = d_x$ units of mass to each square in row x, for a total row-mass of $w'_x = d_x\lfloor B^{k_2-k_1+1}\rfloor$. The rows in $\mathcal{X}_k$ are given exactly k squares, with $s(x, i) = d_x$ units of mass per square; i.e., there is no change for these rows.

It is clear that $\mathcal{S}'$ satisfies (i) by construction. Let us prove that (ii) is also satisfied. First notice that the shape is exactly the same for the rows with $\leq 3(1 + \log T)$ squares. For the remaining rows, we can use the inequalities:

$$B^{k_2-k_1-1} \leq v_x = \frac{w_x}{d_x} < B^{k_2-k_1+1}. \tag{5}$$

From the left-hand side we derive

$$w'_x \le B^2 w_x + d_x < \left(B^2 + \frac{1}{2(1+\log T)}\right) w_x,$$

i.e., the row-masses of $\mathcal{S}'$ are at most C-times larger than the corresponding row-masses of $\mathcal{S}$, with

$$C = B^2 + \frac{1}{2\log T} \le 1 + \frac{1}{1+\log T}.$$

On the other hand, from the right-hand side of (5) we find that $w_x \le w'_x$.

So here is a path from $\mathcal{S}'$ to $\mathcal{S}$ along the edges of $\mathcal{G}$: we go to each row x of $\mathcal{S}'$, and we remove units of mass from the last square of the row until $w'_x = w_x$. How many units must be moved? Because $w'_x \le C w_x$, in row x we only need to remove at most $\frac{1}{1+\log T} w_x$ units. Summing over all the rows, we remove a total of no more than $\frac{T}{1+\log T}$ units, so this upper bounds the length of this path from $\mathcal{S}'$ to $\mathcal{S}$. The item (ii) now follows from the fact that $T' \le B^2 T$, and hence $\log T' \le 1 + \log T$. □