Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 129 (2015)

On the entropy of a noisy function

Alex Samorodnitsky[*]

Abstract

Let $0 < \epsilon < 1/2$ be a noise parameter, and let T_ϵ be the noise operator acting on functions on the boolean cube $\{0,1\}^n$. Let f be a nonnegative function on $\{0,1\}^n$. We upper bound the entropy of $T_\epsilon f$ by the average entropy of conditional expectations of f, given sets of roughly $(1-2\epsilon)^2 \cdot n$ variables.

In information-theoretic terms, we prove the following strengthening of "Mrs. Gerber's lemma": Let X be a random binary vector of length n, and let Z be a noise vector, corresponding to a binary symmetric channel with crossover probability ϵ. Then, setting $v = (1-2\epsilon)^2 \cdot n$, we have (up to lower-order terms):

$$H\Big(X \oplus Z\Big) \quad \geq \quad n \cdot H_2 \left(\epsilon \ + \ (1-2\epsilon) \cdot H_2^{-1} \left(\frac{\mathbb{E}_{|B|=v} H\Big(\{X_i\}_{i \in B}\Big)}{v}\right)\right)$$

Assuming $\epsilon \geq 1/2 - \delta$, for some absolute constant $\delta > 0$, this inequality, combined with a strong version of a theorem of Friedgut, Kalai, and Naor, due to Jendrej, Oleszkiewicz, and Wojtaszczyk, shows that if a boolean function f is close to a characteristic function g of a subcube of dimension $n-1$, then the entropy of $T_\epsilon f$ is at most that of $T_\epsilon g$.

Taken together with a recent result of Ordentlich, Shayevitz, and Weinstein, this shows that the "Most informative boolean function" conjecture of Courtade and Kumar holds for high noise $\epsilon \geq 1/2 - \delta$.

Namely, if X is uniformly distributed in $\{0,1\}^n$ and Y is obtained by flipping each coordinate of X independently with probability ϵ, then, provided $\epsilon \geq 1/2 - \delta$, for any boolean function f holds $I\left(f(X); Y\right) \leq 1 - H(\epsilon)$.

1 Introduction

This paper is motivated by the following conjecture of Courtade and Kumar [7].

Let (X, Y) be jointly distributed in $\{0,1\}^n$ such that their marginals are uniform and Y is obtained by flipping each coordinate of X independently with probability ϵ. Let H_2 denote the binary entropy function $H_2(x) = -x \log_2 x - (1-x)\log_2(1-x)$. The conjecture of [7] is:

Conjecture 1.1: For all boolean functions $f : \ \{0,1\}^n \rightarrow \{0,1\}$,

$$I\Big(f(X); Y\Big) \quad \leq \quad 1 \ - \ H_2(\epsilon)$$

[*] School of Engineering and Computer Science, The Hebrew University of Jerusalem, Jerusalem 91904, Israel. Research supported by BSF and ISF grants.

ISSN 1433-8092

∎

This inequality holds with equality if f is a characteristic function of a subcube of dimension $n-1$. Hence, the conjecture is that such functions are the "most informative" boolean functions.

Following [9], we express $I(f(X);Y)$ in terms of the 'value of the entropy functional of the image of f under the noise operator' (all notions will be defined shortly). The question then becomes:

Which boolean functions are the "stablest" under the action of the noise operator? That is, for which functions the entropy functional decreases the least under noise.

One can also consider a more general question of how the noise operator affects the entropy of a nonnegative function.

Our main result is that for a nonnegative function f on $\{0,1\}^n$, the entropy of the image of f under the noise operator with noise parameter ϵ is upper bounded by the average entropy of conditional expectations of f, given sets of roughly $(1-2\epsilon)^2 \cdot n$ variables.

As an application, using the recent strengthening [6] of a theorem of [4], we show that for ϵ close to $1/2$ characteristic functions of $(n-1)$-dimensional subcubes are at least as stable under the noise operator as functions which are close to them.

This, in conjunction with [4] and a recent result of [14] which can be used to show that, for high noise levels $\epsilon \sim 1/2$, boolean functions, which are potentially as stable as the characteristic functions of $(n-1)$-dimensional subcubes, have to be close to these functions, implies the validity of Conjecture 1.1 for high noise levels.

1.1 Entropy of nonnegative functions and the noise operator

We introduce some relevant notions.

For a nonnegative function $f:\ \{0,1\}^n \to \mathbb{R}$, we let the *entropy* of f to be defined as

$$Ent\Big(f\Big) \quad = \quad \mathop{\mathbb{E}}_x f(x) \log_2 f(x) \quad - \quad \mathop{\mathbb{E}}_x f(x) \cdot \log_2 \Big(\mathbb{E}\, f(x)\Big)$$

We note for future use that entropy is nonnegative, homogeneous $Ent\Big(\lambda f\Big) = \lambda \cdot Ent\Big(f\Big)$ and convex in f [8].

Given $0 \le \epsilon \le 1/2$, we define the *noise operator* acting on functions on the boolean cube as follows: for $f:\ \{0,1\}^n \to \mathbb{R}$, we let $T_\epsilon f$ at a point x be the expected value of f at y, where y is ϵ-correlated with x. That is,

$$\left(T_\epsilon f\right)(x) \quad = \quad \sum_{y \in \{0,1\}^n} \epsilon^{|y-x|} \cdot (1-\epsilon)^{n-|y-x|} \cdot f(y) \qquad (1)$$

Here $|\cdot|$ denotes the Hamming distance.

Note that $T_\epsilon f$ is a convex combination of shifted copies of f. Hence, convexity of entropy implies that the noise operator decreases entropy. Our goal is to quantify this statement.

1.1.1 Connection between notions

Let f be a nonnegative function on $\{0,1\}^n$. Let X be a random variable on $\{0,1\}^n$ distributed according to $f/\sum f$. Let Z be an independent noise random variable on $\{0,1\}^n$. That is, $Pr\{Z = z\} \; = \; \epsilon^{|z|} \cdot (1-\epsilon)^{n-|z|}$, and X and Z are statistically independent. Then

- $$Ent(f) \quad = \quad \mathbb{E} f \cdot \Big(n \; - \; H(X)\Big)$$
- $$Ent\Big(T_\epsilon f\Big) \quad = \quad \mathbb{E} f \cdot \Big(n \; - \; H\Big(X \oplus Z\Big)\Big)$$

Let now $f : \; \{0,1\}^n \to \{0,1\}$ be a boolean function, let X be uniformly distributed in $\{0,1\}^n$, let Z be an independent noise random variable, and let $Y = X \oplus Z$. Then

$$H\Big(f(X)\Big) \quad = \quad Ent\Big(f\Big) \quad + \quad Ent\Big(1-f\Big)$$

We also have the following simple claim (proved in Section 6 below)

Lemma 1.2: *For a boolean function* $f : \; \{0,1\}^n \to \{0,1\}$,

$$I\Big(f(X);Y\Big) \quad = \quad Ent\Big(T_\epsilon f\Big) \quad + \quad Ent\Big(T_\epsilon(1-f)\Big)$$

Therefore, Conjecture 1.1 translates as follows:

Conjecture 1.3: (An equivalent form of Conjecture 1.1)

For any boolean function $f : \; \{0,1\}^n \to \{0,1\}$ holds

$$Ent\Big(T_\epsilon f\Big) \quad + \quad Ent\Big(T_\epsilon(1-f)\Big) \quad \le \quad 1 \; - \; H_2(\epsilon)$$

▮

1.2 Mrs. Gerber's function and Mrs. Gerber's lemma

We describe a result from information theory, and a related function, which will be important for us [1].

Let f_t be a function on the two-point space $\{0,1\}$, which is t at zero and $2-t$ at one. We have

$$Ent\Big(f_t\Big) \quad = \quad 1 \quad - \quad H_2\left(\frac{t}{2}\right)$$

[1] We are grateful to V. Chandar [3] for explaining the relevance of this result in connection to our previous work [18] on the subject.

Let $\phi(x, \epsilon)$ be a function on $[0, 1] \times [0, 1/2]$ defined as follows:

$$\phi(x, \epsilon) \quad = \quad Ent\Big(T_\epsilon f_t\Big) \tag{2}$$

where t is chosen so that $Ent\left(f_t\right) = x$.

This function was introduced in [21]. We will now describe some of its properties.

Note that ϕ is increasing in x, starting from zero at $x = 0$.

In fact, it is easy to derive the following explicit expression for ϕ:

$$\phi(x, \epsilon) \quad = \quad 1 \;-\; H_2\Big((1 - 2\epsilon) \cdot H_2^{-1}(1 - x) \;+\; \epsilon\Big)$$

A key property of ϕ is its concavity.

Theorem 1.4: *([21]) The function $\phi(x, \epsilon)$ is concave in x for any $0 \le \epsilon \le 1/2$.*

We mention a simple corollary.

Corollary 1.5: *For all $0 \le \epsilon \le 1/2$,*

$$\Big(1 - H_2(\epsilon)\Big) \cdot x \quad \le \quad \phi(x, \epsilon) \quad \le \quad (1 - 2\epsilon)^2 \cdot x \tag{3}$$

Proof: It's easy to check $\phi(0, \epsilon) = 0$ and $\phi(1, \epsilon) = 1 - H_2(\epsilon)$. And, it's easy to check that $\frac{\partial \phi}{\partial x}$ at $x = 0$ is $(1 - 2\epsilon)^2$. ▮

From now on, when the value of ϵ is clear from the context, we omit the second parameter in ϕ and write $\phi(x)$ instead of $\phi(x, \epsilon)$.

We now describe an inequality of [21], which is known as Mrs. Gerber's lemma. Following this usage, we will refer to the function ϕ as Mrs. Gerber's function.

This inequality upperbounds the entropy of the image of a nonnegative function under the action of the noise operator. We present it in terms of the entropy functional and the noise operator[2].

Theorem 1.6: *([21]) Let f be a nonnegative function on $\{0, 1\}^n$. Then*

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad n\,\mathbb{E}\,f \cdot \phi\left(\frac{Ent(f)}{n\,\mathbb{E}\,f},\ \epsilon\right) \tag{4}$$

[2] As pointed out to us by Chandar [3], this is equivalent to the standard information-theoretic formulation: Let X be a random binary vector of length n distributed according to $f/\sum f$, and let Z be a noise vector, corresponding to a binary symmetric channel with crossover probability ϵ. Then $H\left(X \oplus Z\right) \ge nH_2\left(\epsilon + (1 - 2\epsilon) \cdot H_2^{-1}\left(\frac{H(X)}{n}\right)\right)$.

1.3 Main results

For $A \subseteq [n]$ and for a nonnegative function $f: \{0,1\}^n \rightarrow \mathbb{R}$, we denote

$$\mathbb{E}\Big(f \mid A\Big) \quad = \quad \mathbb{E}\Big(f \;\Big|\; \{x_i\}_{i \in A}\Big)$$

Here $\mathbb{E}$ is the conditional expectation operator. That is, $\mathbb{E}\Big(f \mid A\Big)$ is the function of the variables $\{x_i\}_{i \in A}$, defined as the expectation of f given the values of $\{x_i\}$.[3]

We write

$$Ent\Big(f \mid A\Big) \quad = \quad Ent\Big(\mathbb{E}\Big(f \mid A\Big)\Big)$$

To connect notions, observe that if X is a random variable on $\{0,1\}^n$ distributed according to $f / \sum f$, then the distribution of $\{X_i\}_{i \in A}$ on the $|A|$-dimensional cube is given by $\frac{1}{2^{|A|}\,\mathbb{E} f} \cdot \mathbb{E}\left(f | A\right)$ and that

$$Ent\Big(f \mid A\Big) \quad = \quad \mathbb{E} f \cdot \Big(|A| \; - \; H\Big(\{X_i\}_{i \in A}\Big)\Big) \tag{5}$$

Our main claim is that the entropy of a nonnegative function f under noise is upper bounded by the average entropy of conditional expectations of f, given certain random subsets of variables. We present several results which illustrate this fact.

Theorem 1.7: *Let f be a nonnegative function on the cube with $\mathbb{E} f = 1$.*

Let $0 < \epsilon < 1$ be a noise parameter. Let T be a random subset of $[n]$ generated by sampling each element $i \in [n]$ independently with probability $(1-2\epsilon)^2$. Then

$$Ent\Big(T_\epsilon f\Big) \quad \leq \quad \mathop{\mathbb{E}}_{T} \quad \Big(Ent\Big(f \mid T\Big) \; - \; \sum_{i \in T} Ent\Big(f \mid \{i\}\Big)\Big) \quad + \quad \sum_{i=1}^{n} \phi\Big(Ent\Big(f \;\Big|\; \{i\}\Big)\Big)$$

Remark 1.8: We are grateful to O. Ordentlich for suggesting this formulation for the claim of this theorem, as well as for Theorem 1.12 below (in earlier versions the average on the RHS was taken over sets of a fixed cardinality $\sim (1-2\epsilon)^2 \cdot n$, which led to more cumbersome calculations.)

Let us also mention that Polyanskiy and Wu [17] came up with a new and direct proof of the key claim, Proposition 4.1, which does not rely on linear programming, and this was used by Ordentlich [12] to give direct proofs for Theorems 1.7 and 1.12. ▮

Applying the inequality $\phi(x,\epsilon) \leq (1-2\epsilon)^2 \cdot x$ (see (3)) to the claim of the theorem, gives the following, more streamlined claim. (However, the somewhat stronger claim of the theorem is needed for the applications.)

[3] We also may (and will) view $\mathbb{E}\left(f \mid A\right)$ as a function on $\{0,1\}^n$, which depends only on variables with indices in A.

Corollary 1.9: *In the notation of Theorem 1.7,*

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad \mathop{\mathbb{E}}_{T} Ent\Big(f \mid T\Big)$$

Specializing to boolean functions, this implies the following claim.

Corollary 1.10: *In the notation of Conjecture 1.1 and of Theorem 1.7, for a boolean function* $f:\ \{0,1\}^n \rightarrow \{0,1\}$ *holds*

$$I\Big(f(X);Y\Big) \quad \le \quad \mathop{\mathbb{E}}_{T}\ I\Big(f(X);\ \{X_i\}_{i\in T}\Big)$$

Remark 1.11: Let B be a random subset of $[n]$ generated by sampling each element $i \in [n]$ independently with probability $1-2\epsilon$.

As pointed out by Or Ordentlich [12], it seems instructive to compare the bound in Corollary 1.10 to the weaker bound

$$I\Big(f(X);Y\Big) \quad \le \quad \mathop{\mathbb{E}}_{B}\ I\Big(f(X);\ \{X_i\}_{i\in B}\Big)$$

which can be obtained by the following information-theoretic argument.

An equivalent way to obtain Y from X is to replace each coordinate of X independently with a random bit, with probability 2ϵ .

Let S be the set of indices where the input bits were replaced with random bits, and let $B = S^c$.

Using the chain rule of mutual information we have

$$I\Big(f(X);Y\Big) \quad = \quad I\Big(f(X);Y,S\Big) \ - \ I\Big(f(X);S \mid Y\Big) \quad = \quad I\Big(f(X);Y \mid S\Big) \ - \ I\Big(f(X);S \mid Y\Big)$$

where the last equality follows since $I\Big(f(X);S\Big) = 0$.

In particular, by non-negativity of mutual information

$$I\Big(f(X);Y\Big) \quad \le \quad I\Big(f(X);Y \mid S\Big) \quad = \quad \mathbb{E}_B\ I\Big(f(X);\{X_i\}_{i\in B}\Big)$$

▮

We also show a somewhat different strengthening of Corollary 1.9, which gives a stronger version of Mrs. Gerber's lemma (Theorem 1.6).

Theorem 1.12: *In the notation of Theorem 1.7, setting* $t = (1-2\epsilon)^2 \cdot n$*, the following is true:*

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad n \cdot \phi\left(\frac{\mathbb{E}_T\, Ent\Big(f \mid T\Big)}{t}\ ,\ \ \epsilon\right)$$

In the standard information-theoretic notation, this could be restated as follows. Let X be a random binary vector of length n, and let Z be an independent noise vector, corresponding to a binary symmetric channel with crossover probability ϵ. Then

$$H\Big(X \oplus Z\Big) \quad \geq \quad n \cdot H_2 \left(\epsilon + (1-2\epsilon) \cdot H_2^{-1} \left(\frac{\mathbb{E}_T \, H\Big(\{X_i\}_{i \in T}\Big)}{t} \right) \right) \tag{6}$$

We refer to [13] for an application of (6).

Remark 1.13:

Up to a negligible error term, the claim of the theorem is stronger than that of Theorem 1.6, since the sequence $a_t = \frac{\mathbb{E}_{|T|=t} H(\{X_i\}_{i \in T})}{t}$ is increasing, by Han's inequality [5].

▮

We now return to Conjectures 1.1 and 1.3.

Let us first describe a family of functions for which these conjectures are known to hold with equality. Let $1 \leq k \leq n$ be an index, and let $g_k(x) = 1$ if and only if $x_k = 0$. (That is, g_k is a characteristic function of the $(n-1)$-dimensional subcube $\{x_k = 0\}$.)

It is easy to verify that $Ent\,(T_\epsilon g_k) = \frac{1}{2} \cdot (1 - H_2(\epsilon))$ and $Ent\,(T_\epsilon g_k) + Ent\,(T_\epsilon\,(1 - g_k)) = 1 - H_2(\epsilon)$.

We apply Theorem 1.7 to show that, for $\epsilon \sim 1/2$, the conjectures also hold for functions which are close to characteristic functions of subcubes.

To make the notion of proximity more precise, recall (see [11]) that any function $f : \{0,1\}^n \rightarrow \mathbb{R}$ can be expanded in terms of the Walsh-Fourier basis: $f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \cdot W_S(x)$. Here $W_S(x) = (-1)^{\sum_{i \in S} x_i}$.

The Walsh-Fourier expansion of g_k is especially simple: $\widehat{g_k}(0) = \mathbb{E}\, g_k = 1/2$, $\widehat{g_k}(\{k\}) = 1/2$, and $\widehat{g_k}(S) = 0$ for all other $S \subseteq [n]$.

It follows from [6] and [4] that a boolean function whose Walsh-Fourier expansion is close to that of g_k, in that it has a large (i.e., close to 1/2) Fourier coefficient at $\{k\}$, has to be very close, in the appropriate sense, to g_k.

The next claim shows the conjectures to hold for such functions.

Theorem 1.14: *There exists an absolute constant $\delta > 0$ such that for any noise $\epsilon \geq 0$ with $(1-2\epsilon)^2 \leq \delta$ and for any boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$ such that*

- $\frac{1}{2} - \delta \leq \mathbb{E}\, f \leq \frac{1}{2}$*;*
- *There exists $1 \leq k \leq n$ such that $|\widehat{f}(\{k\})| \geq (1-\delta) \cdot \mathbb{E}\, f$*

Holds

1. $$Ent\Big(T_\epsilon f\Big) \quad \le \quad \frac{1}{2}\cdot\Big(1-H_2(\epsilon)\Big)$$

2. $$Ent\Big(T_\epsilon f\Big) \; + \; Ent\Big(T_\epsilon(1-f)\Big) \quad \le \quad 1-H_2(\epsilon)$$

This, in conjunction with [4] and [14], which can be used to show that, for noise parameter close to $1/2$, boolean functions, which are potentially as stable as the characteristic functions of $(n-1)$-dimensional subcubes, have to satisfy the constraints of Theorem 1.14, implies the validity of Conjecture 1.1 for high noise levels.

Theorem 1.15: *There exists an absolute constant $\delta > 0$ such that for any noise $\epsilon \ge 0$ with $(1-2\epsilon)^2 \le \delta$ and for any boolean function $f:\ \{0,1\}^n \to \{0,1\}$ holds*

$$I\Big(f(X);Y\Big) \quad \le \quad 1 \; - \; H_2(\epsilon)$$

1.4 More on Theorems 1.7 and 1.12

In this subsection we give a high-level description of the proofs of these theorems and argue that both their claims may be viewed as strengthenings of Mrs. Gerber's lemma.

Notation: For a direction $1 \le i \le n$ we define the noise operator in direction i as follows:

$$\Big(T_{\epsilon_{\{i\}}} f\Big)(x) \quad = \quad \epsilon\cdot f\Big(x+e_i\Big) \quad + \quad (1-\epsilon)\cdot f(x)$$

where e_i is the i^{th} unit vector. The operators $\Big\{T_{\epsilon_{\{i\}}}\Big\}$ commute and, for $R \subseteq [n]$, we define T_{ϵ_R} to be the composition of $T_{\epsilon_{\{i\}}}$, $i \in R$. Note that the noise operator T_ϵ would be written in this notation as $T_{\epsilon_{[n]}}$.

We start with the proof of Mrs. Gerber's lemma (4). Since both sides of the inequality are homogeneous in f, we may assume $\mathbb{E} f = 1$.

By the chain rule for entropy, for any permutation σ in the symmetric group S_n holds

$$Ent\Big(T_\epsilon f\Big) \quad = \quad \sum_{i=1}^n \left(Ent\Big(T_\epsilon f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) \; - \; Ent\Big(T_\epsilon f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\right) \quad =$$

$$\sum_{i=1}^n \left(Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) \; - \; Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\right) \quad \le$$

$$\sum_{i=1}^n \phi\left(Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\right) \tag{7}$$

Let us explain the last inequality. Let $y \in \{0,1\}^{i-1}$. Let $\tilde{f}_y$ be a function on $\{0,1\}$ defined by the restriction of the function $\mathbb{E}\left(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\right)$, which we view as a function on the i-dimensional cube, to the points in which the coordinates $\sigma(k)$, $k = 1, ..., i-1$ are set to be y_k. Then, it is easy to see that

$$Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i)\}}} f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big) =$$

$$\mathop{\mathbb{E}}_{y} Ent\Big(T_\epsilon \tilde{f}_y\Big) = \mathop{\mathbb{E}}_{y}\left(\mathbb{E}\,\tilde{f}_y \cdot \phi\left(Ent\left(\frac{\tilde{f}_y}{\mathbb{E}\,\tilde{f}_y}\right)\right)\right) \leq \phi\Big(\mathop{\mathbb{E}}_{y} Ent\Big(\tilde{f}_y\Big)\Big) =$$

$$\phi\Big(Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\Big)$$

The first equality in the second row follows from (2) and the linearity of entropy. The inequality follows from concavity of the function ϕ and the fact that $\mathbb{E}_y \mathbb{E}\,\tilde{f}_y = \mathbb{E}\left(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\right)$ $= \mathbb{E}\,f = 1$.

We now continue from (7).

For $y \in \{0,1\}^{i-1}$, let f_y be a function on $\{0,1\}$ defined by the restriction of the function $\mathbb{E}\left(f \mid \{\sigma(1),\ldots,\sigma(i)\}\right)$ to the points in which the coordinates $\sigma(k)$, $k = 1, ..., i-1$ are set to be y_k.

Since the noise operator $T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}}$ is stochastic, the functions $\left\{\tilde{f}_y\right\}$ are a stochastic mixture of the functions $\left\{f_y\right\}$. Hence, since the *Ent* functional is convex, for any $0 \leq \epsilon \leq 1$ holds

$$Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big) =$$

$$\mathop{\mathbb{E}}_{y} Ent\Big(\tilde{f}_y\Big) \leq \mathop{\mathbb{E}}_{y} Ent\Big(f_y\Big) = \qquad (8)$$

$$Ent\Big(f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)$$

And hence (7) is upper bounded by

$$\sum_{i=1}^{n} \phi\Big(Ent\Big(f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) - Ent\Big(f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\Big) \leq n \cdot \phi\left(\frac{Ent\Big(f\Big)}{n}\right)$$

where in the last inequality the concavity of ϕ is used again.

1.4.1 Our improvement

We attempt to quantify the loss in inequality (8).

Let us introduce some notation. For a nonnegative function g on the cube, for a subset $A \subset [n]$, and for an element $m \notin A$, we define

$$I_g(A,m) \quad = \quad Ent\Big(g \mid A \cup \{m\}\Big) \quad - \quad Ent\Big(g \mid A\Big) \quad - \quad Ent\Big(g \mid \{m\}\Big)$$

This quantity is always nonnegative. In fact, let X be distributed on $\{0,1\}^n$ according to $g/\sum g$. Assume $\mathbb{E}\, g = 1$ and note that in this case, by Subsection 1.1.1 and by (5), we have $I_g(A,m) \ = \ H\left(\{X_i\}_{i \in A}\right) + H\left(X_m\right) - H\left(\{X_j\}_{j \in A \cup \{m\}}\right) \ = \ I\left(\{X_i\}_{i \in A};\ X_m\right)$.

Coming back to (8), observe that $Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f \mid \{\sigma(i)\}\Big) \ = \ Ent\Big(f \mid \{\sigma(i)\}\Big)$.

Hence, taking $A = \{\sigma(1),\ldots,\sigma(i-1)\}$ and $m = \{\sigma(i)\}$, the decrease in (8) is from $I_f(A,m)$ to $I_{T_{\epsilon_A} f}(A,m)$. Therefore, our goal is to quantify the decrease in mutual information in the presence of noise.

In the next two sections we consider a somewhat more general question of upper bounding $I_{T_{\epsilon_A} f}(A,m)$, given f, A, and m. In Section 2 we upper bound $I_{T_{\epsilon_A} f}(A,m)$ by the value of a certain linear program. In Section 3 we introduce a symmetric version of this program and a symmetric solution for the symmetric program, and show its value to be at least as large as that of the original program.

We then find the value of the symmetric solution, as a function of f, A, and m. This value provides an upper bound on the noisy mutual information (see Proposition 4.1).

In order to prove Theorems 1.7 and 1.12 we apply the improved bound in (8), averaging the chain rule for the entropy of $T_\epsilon f$ over all permutations $\sigma \in S_n$.

This improvement in (8) is the reason we suggest to view both these claims as stronger versions of Mrs. Gerber's lemma.

On the other hand, strictly speaking, this line of argument does not necessarily provide a direct improvement of (4), since in the averaging step we have to replace $\phi(x,\epsilon)$ by a larger linear function $(1-2\epsilon)^2 \cdot x$, in order to be able to come up with manageable estimates.

In fact, the difference between the two claims stems from the different ways in which we apply this "linearization" of the function $\phi(x,\epsilon)$ during averaging. The bounds they give are incomparable, though Theorem 1.12 is a more evident improvement of (4).

We note that the two functions $\phi(x,\epsilon)$ and $(1-2\epsilon)^2 \cdot x$ almost coincide for small values of x, and, loosely speaking, if the entropy of f is not too large, as is the case, say, for boolean functions, all the arguments of ϕ should lie very close to zero, meaning not much lost in the linear approximation. In this case, the bounds in Theorems 1.7 and 1.12 are very close to that in Corollary 1.9.

1.4.2 Related work

Y. Polyanskiy [15] has pointed out to us that the related question of upper bounding $I_{T_{\epsilon_A} f}(A, m)$ given $I_f(A, m)$ belongs to the area of *strong data processing inequalities* (SDPI) in information theory (see [16], [17] for pertinent results, and, in particular, for a new proof of Proposition 4.1).

Organization of the paper

This paper is organized as follows. The proof of Theorem 1.7 is given in Sections 2 to 4. Theorem 1.14 is proved in Section 5. The remaining proofs are presented in Section 6.

2 A linear programming bound for noisy mutual information

In this section we upper bound the noisy mutual information $I_{T_{\epsilon_A} f}(A, m)$ by the value of a certain linear program.

Let f be a nonnegative function on the cube. Let A be a subset of $[n]$ and let $m \not\in A$.

Let $|A| = k$. We will assume, without loss of generality, that $A = [k]$ and that $m = k + 1$.

Notation: From now on, we write λ for $(1 - 2\epsilon)^2$.

Discussion

Before going into details, let us give a high-level description of what the linear program attempts to capture. For ease of discussion the notation we use here is slightly different from that in the definition of the program below (they are the same up to scaling).

Given a random variable X on $\{0, 1\}^n$ distributed according to $f / \sum f$, consider a function I on the k-dimensional boolean cube, defined for $S \subseteq [k]$ by the mutual information $I(S) = I(\{X_i\}_{i \in S};\ X_{k+1})$.

For $S \subseteq [k]$ and for $i \in S$, let $y_{S,i} = I(S) - I(S \setminus \{i\})$ be the "discrete derivative" of I at S in direction i. Note that $y_{S,i} \geq 0$, since this is the mutual information between X_i and X_{k+1}, given $\{X_j\}_{j \in S \setminus \{i\}}$. We view y as a function on the edges of the cube. Note also that, for any S, the value of the summation of y on the edges of any path from $\emptyset$ to S is $I(S)$.

For $R \subseteq [k]$, applying noise in directions in R to f leads to a new distribution $T_{\epsilon_R} f / \left(\sum T_{\epsilon_R} f\right)$ on $\{0, 1\}^n$. This defines a new random variable X^R, a mutual information function I^R and discrete derivative functions $x^R_{S,i} = I^R(S) - I^R(S \setminus \{i\})$. (Note that $x^\emptyset = y$).

Observe that noise decreases mutual information, and hence $I^R \leq I$. However, the discrete derivatives x^R do not necessarily decrease. With that, and this is a key fact, by the *strong data processing inequality* [2], noise in direction i decreases the discrete derivative in direction i (i.e., the conditional mutual information between X^R_i and X^R_{k+1}) by a factor of at least λ.

The variables in the linear program below are the values of the discrete derivates x^R, while we consider the discrete derivatives $y = x^\emptyset$ related to the initial function f to be the boundary data of the program. We note that the noisy mutual information $I([k]) = I_{T_{\epsilon_{[k]}} f}([k], k+1)$ is a linear combination of the variables, and that the strong data processing inequality provides linear local constraints on the variables.

Finally, we would like to explain the intuition behind the symmetrization procedure in Section 3. The fact that for any R and S the value of the summation of x^R on the edges of any path from $\emptyset$ to S is $I^R(S)$ provides a family of "symmetric" linear constraints on the variables. This makes it natural to look for a symmetric feasible solution to the linear program (symmetrizing the boundary data accordingly), one in which $x^R(S,i)$ depends only on $|S|$ and on $|R \cap S|$.

We were led to expect that this symmetric solution would be an optimal one by the following informal speculation. It turns out that the strong data processing inequality $x^R_{S,i} \le \lambda \cdot \left(x^{R\setminus i}_{S,i}\right)$ may be replaced by a stronger inequality $x^R_{S,i} \le \phi\left(x^{R\setminus i}_{S,i}\right)$ (see (3)).[4] This turns the program into a *strictly concave* optimization problem, for which optimality of a feasible symmetric solution might be anticipated. It might also be hoped for that replacing the concave constraint by a linear one would preserve this property, and this is indeed turns out to be true.

More to the point, it turns out that for the symmetric solution we define, all the inequalities $x^R_{S,i} \le \lambda \cdot \left(x^{R\setminus i}_{S,i}\right)$ hold with equality.

The resulting argument is straightforward, most of the work going into setting up notation, and verifying feasibility of the symmetric solution. The key step, relying on symmetric properties of the discrete cube, is made in Lemma 3.6.

Linear program

Boundary data: For $S \subseteq [k]$ and for $i \in S$, we write

$$y_{S,i} \quad = \quad Ent\Big(f \mid S \cup \{k+1\}\Big) \; - \; Ent\Big(f \mid S \setminus \{i\} \cup \{k+1\}\Big) \; - \; Ent\Big(f \mid S\Big) \; + \; Ent\Big(f \mid S \setminus \{i\}\Big)$$

The numbers $\{y_{S,i}\}$ are the boundary data for this problem.

We note that $y_{S,i} \ge 0$ for all S and i. In fact, the value of $y_{S,i}$ is proportional to a certain conditional mutual information. To see this, let X be distributed on $\{0,1\}^n$ according to $f/\sum f$. Assume $\mathbb{E} f = 1$ and note that, by Subsection 1.1.1 and by (5), $y_{S,i}$ is given by
$H\left(\{X_i\}_{i\in S\setminus\{i\}\cup\{k+1\}}\right) + H\left(\{X_i\}_{i\in S}\right) - H\left(\{X_i\}_{i\in S\cup\{k+1\}}\right) - H\left(\{X_i\}_{i\in S\setminus\{i\}}\right) \; = \; I\left(X_i; X_{k+1} | \{X_j\}_{j\in S\setminus\{i\}}\right).$

Variables: $x^R_{S,i}$ for $R, S \subseteq [k]$ and $i \in S$.

The optimization problem: Given the boundary data, we want to upper bound μ, where

$$\mu \quad = \quad \text{Max} \quad \sum_{i=1}^{k} x^{[k]}_{\{1,\ldots,i\};\; i} \tag{9}$$

under the following constraints.

Constraints:

[4] This was shown in [19] if f is monotone (which suffices for applications) and in [17] for general functions.

1. $$x^{\emptyset}_{S,i} \quad = \quad y_{S,i}$$

2. $$x^{R}_{S,i} \quad = \quad x^{R \cap S}_{S,i}$$

3. For all $\sigma, \tau \in S_k$ holds

$$\sum_{i=1}^{k} x^{R}_{\{\sigma(1),\ldots,\sigma(i)\},\ \sigma(i)} \quad = \quad \sum_{i=1}^{k} x^{R}_{\{\tau(1),\ldots,\tau(i)\},\ \tau(i)}$$

4. If $i \in R$ then

$$x^{R}_{S,i} \quad \le \quad \lambda \cdot \left(x^{R \setminus i}_{S,i} \right)$$

We then have the following claim.

Theorem 2.1: *The noisy mutual information $I_{T_{\epsilon_{[k]}} f}\Big([k], k+1\Big)$ is upperbounded by the value of the optimization problem (9).*

Proof:

First, consider the boundary data. We claim that for any permutation $\sigma \in S_k$ holds

$$\sum_{i=1}^{k} y_{\{\sigma(1),\ldots,\sigma(i)\},\ \sigma(i)} \quad = \quad I_f\Big([k], k+1\Big) \tag{10}$$

In fact, it is easy to see that the LHS is a telescopic sum, summing to

$$Ent\Big(f \mid [k+1]\Big) \quad - \quad Ent\Big(f \mid [k]\Big) \quad - \quad Ent\Big(f \mid \{k+1\}\Big) \quad = \quad I_f\Big([k], k+1\Big)$$

Next we define a feasible solution for (9) whose value is $I_{T_{\epsilon_{[k]}} f}\Big([k], k+1\Big)$.

Fix $R \subseteq [k]$. Write f^R for $T_{\epsilon_R} f$. For $S \subseteq [k]$ and $i \in S$ set

$$x^{R}_{S,i} \;=\; Ent\Big(f^R \mid S \cup \{k+1\}\Big) - Ent\Big(f^R \mid S \setminus \{i\} \cup \{k+1\}\Big) - Ent\Big(f^R \mid S\Big) + Ent\Big(f^R \mid S \setminus \{i\}\Big)$$

Clearly, $x^{\emptyset}_{S,i} = y_{S,i}$ and hence the first constraint of the program is satisfied.

As above, for any permutation $\sigma \in S_k$ holds

$$\sum_{i=1}^{k} x^{R}_{\{\sigma(1),\ldots,\sigma(i)\},\ \sigma(i)} \quad = \quad I_{T_{\epsilon_R}} f\Big([k], k+1\Big)$$

Hence, the third constraint is satisfied as well.

In particular,

$$\sum_{i=1}^{k} x_{\{1,\ldots,i\},\ i}^{[k]} \quad = \quad I_{T_{\epsilon_{[k]}}} f\Big([k], k+1\Big)$$

so, the value given by this solution is indeed $I_{T_{\epsilon_{[k]}}} f\Big([k], k+1\Big)$.

We continue to prove its feasibility. We claim that for any $A \subseteq [k]$ holds $Ent\Big(f^R \mid A\Big) \;=\; Ent\Big(f^{R\cap A} \mid A\Big)$.

To see this, note that the noise operators commute with the conditional expectation operators, and hence

$$\mathbb{E}\Big(T_{\epsilon_R} f \mid A\Big) \;=\; T_{\epsilon_R}\,\mathbb{E}\Big(f \mid A\Big) \;=\; T_{\epsilon_{R\cap A}} T_{\epsilon_{R\setminus A}}\,\mathbb{E}\Big(f \mid A\Big) \;=\; T_{\epsilon_{R\cap A}}\,\mathbb{E}\Big(f \mid A\Big) \;=\; \mathbb{E}\Big(T_{\epsilon_{R\cap A}} f \mid A\Big)$$

Hence, by definition, $x_{S,i}^R = x_{S,i}^{R\cap S}$ for any $R, S \subseteq [k]$, and the second constraint holds.

To conclude the proof of the theorem, it remains to show that for any $R \subseteq S \subseteq [k]$ and $i \in R$ holds

$$x_{S,i}^R \quad \le \quad \lambda \cdot \Big(x_{S,i}^{R\setminus i}\Big) \tag{11}$$

Recall that the strong data processing inequality [2] for a binary symmetric channel with crossover probability ϵ states that if V is a random variable with values in $\{0,1\}$, and U is any random variable; and if $Y = V \oplus Z$, where Z is a Bernoulli random variable with parameter ϵ, statistically independent of U and V, then $I(U;Y) \le \lambda \cdot I(U;V)$.

Let X be distributed on $\{0,1\}^n$ according to $f^{R\setminus\{i\}} / \sum f^{R\setminus\{i\}}$. Assuming, as we may, $\mathbb{E} f = \mathbb{E} f^{R\setminus\{i\}} = 1$, we can rewrite (11) as

$$I\left(X_i \oplus Z\ ; X_{k+1} \;\middle|\; \Big\{X_j\Big\}_{j\in S\setminus\{i\}}\right) \quad \le \quad \lambda \cdot I\left(X_i\ ; X_{k+1} \;\middle|\; \Big\{X_j\Big\}_{j\in S\setminus\{i\}}\right)$$

which follows from applying the strong data processing inequality with $U = X_{k+1}$ and $V = X_i$, both conditioned on $\{X_j = x_j\}_{j\in S\setminus\{i\}}$, for all values of x_j.

3 The optimization problem and its symmetric version

In this section we introduce a symmetric version of the optimization problem (9) and a specific symmetric feasible solution for the symmetric problem. We then argue that the value of this solution for the symmetric problem is at least as large as the optimal value for the original problem. Hence this value provides an upper bound on the noisy mutual information.

3.1 The symmetric problem and solution

Let $\left\{x_{S,i}^R\right\}$ be a feasible solution to the optimization problem (9) with boundary data $\{y_{S,i}\}$. We define numbers $y_1, \ldots, y_k$ as follows. For $1 \le s \le k$ let

$$y_s \quad = \quad \mathop{\mathbb{E}}_{(S,i)} y_{S,i} \tag{12}$$

where the expectation is taken over all pairs (S, i) such that $|S| = s$ and $i \in S$.

For $0 \le r < s \le k$ we define x_s^r recursively in the following manner:

$$x_s^r \quad = \quad \begin{cases} y_s & \text{if} \qquad r = 0 \\ \lambda \cdot x_s^{r-1} \ + \ (1-\lambda) \cdot x_{s-1}^{r-1} & \text{otherwise} \end{cases} \tag{13}$$

We now define the *symmetric version* of (9), by replacing the boundary data by a new, symmetric one. We set, for all $i \in S \subseteq [k]$ with $|S| = s$:

$$\bar{y}_{S,i} \quad = \quad y_s$$

Next, we define the *symmetric solution* for the symmetric problem, in the following way. For $R \subseteq S$ with $|R| = r$, we set

$$\bar{x}_{S,i}^R \quad = \quad \begin{cases} \lambda \cdot x_s^{r-1} & \text{if} \qquad i \in R \\ x_s^r & \text{otherwise} \end{cases}$$

and for general R, S we set

$$\bar{x}_{S,i}^R \quad = \quad \bar{x}_{S,i}^{R \cap S}$$

Proposition 3.1: *The solution above is a feasible solution of the symmetric version of (9). Moreover, for any $R \subseteq [k]$ of cardinality r and for any $\tau \in S_k$ holds*

$$\sum_{i=1}^k \bar{x}_{\{\tau(1),\ldots,\tau(i)\},\tau(i)}^R \quad = \quad \sum_{j=1}^{k-r} y_j \quad + \quad \lambda \cdot \sum_{t=0}^{r-1} x_{k-r+t+1}^t \tag{14}$$

Proof:

The constraints 1 and 2 of (9) hold, by the definition of $\bar{x}_{S,i}^R$. We pass to constraint 4. Clearly, because of constraint 2, it suffices to prove it for $R \subseteq S$. In this case, taking $i \in R$, we have, by the definition of $\bar{x}_{S,i}^R$

$$\bar{x}_{S,i}^R \quad = \quad \lambda \cdot x_s^{r-1} \quad = \quad \lambda \cdot \bar{x}_{S,i}^{R \setminus \{i\}}$$

Next, we note that (14) will imply validity of constraint 3, since the RHS of (14) does not depend on τ.

It remains to prove (14). Let $i_1 < i_2 < ... < i_r$ be such that $R = \{\tau(i_1), \tau(i_2), ..., \tau(i_r)\}$. Then

$$\sum_{i=1}^{k} \bar{x}^R_{\{\tau(1),...,\tau(i)\},\tau(i)} \quad = \quad \sum_{j=1}^{i_1-1} \quad + \quad \sum_{j=i_1}^{i_2-1} \quad + \quad \dots \quad + \quad \sum_{j=i_r}^{k} \quad =$$

$$\sum_{j=1}^{i_1-1} y_j \quad + \quad \left(\lambda \cdot y_{i_1} + \sum_{j=i_1+1}^{i_2-1} x^1_j\right) \quad + \quad \left(\lambda \cdot x^1_{i_2} + \sum_{j=i_2+1}^{i_3-1} x^2_j\right) \quad + \dots \left(\lambda \cdot x^{r-1}_{i_r} + \sum_{j=i_r+1}^{k} x^r_j\right)$$

Expanding $x^t_s = \lambda \cdot x^{t-1}_s + (1-\lambda) \cdot x^{t-1}_{s-1}$, we have the following exchange rule:

Two adjacent summands of the form $\lambda \cdot x^t_j + x^{t+1}_{j+1}$ can always be replaced by $x^t_j + \lambda \cdot x^t_{j+1}$. Applying this appropriate number of times in each bracket transforms the expression above into

$$\sum_{j=1}^{i_1-1} y_j \quad + \quad \left(\sum_{j=i_1}^{i_2-2} y_j + \lambda \cdot y_{i_2-1}\right) \quad + \quad \left(\sum_{j=i_2}^{i_3-2} x^1_j + \lambda \cdot x^1_{i_3-1}\right) \quad + \dots \left(\sum_{j=i_r}^{k-1} x^{r-1}_j + \lambda \cdot x^{r-1}_k\right)$$

Next we observe that the following rules apply in the original ordering of the summands: To the right of x^t_j is always either x^t_{j+1} or $\lambda \cdot x^t_{j+1}$. To the right of $\lambda \cdot x^r_s$ is always either x^{r+1}_{s+1} or $\lambda \cdot x^{r+1}_{s+1}$.

Moreover, this is easily verified to be preserved by the exchange rule above, by checking the four arising cases.

This means that applying the exchange rule as many times as needed, we can ensure all the summands multiplied by λ to be on the last r places on the right. Since the first summand is always either y_1 or $\lambda \cdot y_1$, these invariants guarantee that by doing so we obtain (14).

∎

3.2 Optimality of the symmetric solution

Theorem 3.2: *Let* $\left\{x^R_{S,i}\right\}$ *be a feasible solution to the linear optimization problem (9). Let* $\left\{\bar{x}^R_{S,i}\right\}$ *be the symmetric solution for the symmetric version of this problem.*

Then, for any $0 \le r \le k$ *holds:*

$$\mathop{\mathbb{E}}_{|R|=r} \sum_{i=1}^{k} x^R_{\{1,...,i\},i} \quad \le \quad \mathop{\mathbb{E}}_{|R|=r} \sum_{i=1}^{k} \bar{x}^R_{\{1,...,i\},i}$$

Corollary 3.3: *The optimal value of (9) is upper bounded by the value of the symmetric solution to the symmetric version of the problem, which is given by*

$$\lambda \cdot \sum_{t=0}^{k-1} x_{t+1}^{t}$$

Proof: Apply the theorem with $r = k$ and use (14). ▮

Proof: (Of the theorem).

We proceed by double induction - on k and on $0 \leq r \leq k$. For $k = 1$ the claim is easily seen to be true.

Note also that the claim is true for any k and $r = 0$. This follows from constraints 1 and 3 of the linear program (9) and the definition of the symmetric boundary data. In fact, we have

$$\sum_{j=1}^{k} y_{\{1,\ldots,j\},j} \quad = \quad \mathop{\mathbb{E}}_{\sigma \in S_k} \sum_{j=1}^{k} y_{\{\sigma(1),\ldots,\sigma(j)\},\ \sigma(j)} \quad = \quad \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\sigma \in S_k} y_{\{\sigma(1),\ldots,\sigma(j)\},\ \sigma(j)} \quad =$$

$$\sum_{j=1}^{k} \mathop{\mathbb{E}}_{|S|=j,\ i \in S} \quad y_{S,i} \quad = \quad \sum_{j=1}^{k} y_j \quad = \quad \sum_{j=1}^{k} \bar{y}_{\{1,\ldots,j\},j}$$

Let now numbers r and k, with $0 < r \leq k$ be given. Assume the claim holds for $k-1$, and also for k, for all $0 \leq t \leq r-1$. We will argue it also holds for k and r.

We start with some simple properties of the linear program (9). We assume to be given the boundary data and a specific feasible solution to (9), and the symmetric solution to the symmetric version of (9), as in Theorem 3.2.

Lemma 3.4: *Let $M \subseteq [k]$. Let $\left\{ y_{K,i} \right\}_{i \in K \subseteq M}$ be the restriction of the boundary data to subsets of M. For $R \subseteq M$, let $\left\{ x_{K,i}^{R} \right\}_{i \in K \subseteq M}$ be the restriction of the feasible solution to subsets of M. Then $\left\{ x_{K,i}^{R} \right\}_{i \in K \subseteq M}$ is a feasible solution to the appropriate (smaller) optimization problem on M.*

Proof:

Constraints 1, 2, and 4 are easy to check. As for constraint 3, let σ and τ be two permutations from M to itself. Extend them in the same way to permutations σ' and τ' on $[k]$. It is then easy to see that constraint 3 holds for σ and τ in the smaller problem, since it holds for σ' and τ' in the larger one.

▮

Lemma 3.5: *Let $M \subseteq [k]$, with $|M| = m$ and let $R \subseteq [k]$. Let τ be a bijection from $[m]$ to M. Let*

$$F\Big(M, R, \tau\Big) \quad = \quad \sum_{j=1}^{m} \bar{x}^{R}_{\{\tau(1),\ldots,\tau(j)\},\ \tau(j)}$$

Then $F\Big(M, R, \tau\Big)$ depends only on m and $|R \cap M|$.

Proof:

Since the symmetric solution $\Big\{\bar{x}^{R}_{S,i}\Big\}$ satisfies constraint 2 of (9), we have

$$F\Big(M, R, \tau\Big) \quad = \quad \sum_{j=1}^{m} \bar{x}^{R}_{\{\tau(1),\ldots,\tau(j)\},\ \tau(j)} \quad = \quad \sum_{j=1}^{m} \bar{x}^{R\cap M}_{\{\tau(1),\ldots,\tau(j)\},\ \tau(j)} \quad = \quad F\Big(M, R \cap M, \tau\Big)$$

Let $r = |R \cap M|$.

Proceeding exactly as in the proof of Proposition 3.1, we get that

$$F\Big(M, R \cap M, \tau\Big) \quad = \quad \sum_{j=1}^{m-r} y_j \quad + \quad \lambda \cdot \sum_{t=0}^{r-1} x^{t}_{m-r+t+1}$$

That is, $F\Big(M, R, \tau\Big)$ depends only on m and $r = |R \cap M|$, as claimed. ∎

Next, we introduce some notation.

3.2.1 Notation

1. Let $M \subseteq [k]$. Let $\Big\{y_{K,i}\Big\}_{i \in K \subseteq M}$ be the restriction of the boundary data to the subsets of M.

   We will denote by $\Big\{\mathcal{S}_M\Big[x^{R}_{K,i}\Big]\Big\}$ the symmetric solution to the symmetric version of the smaller problem with this boundary data.

2. Let $L \subseteq [k]$, with $L = \{i_1, ..., i_\ell\}$, so that $i_1 < i_2 < ... < i_\ell$. Let $R \subseteq [k]$. Write

   $$\mu^{R}(L) \quad = \quad \sum_{j=1}^{\ell} x^{R}_{\{i_1,\ldots,i_j\},\ i_j}$$

   For $L \subseteq M \subseteq [k]$, and $R \subseteq M$, we denote

   $$\mathcal{S}[\mu]^{R}_{M}(L) \quad = \quad \sum_{j=1}^{\ell} \mathcal{S}_M\Big[x^{R}_{\{i_1,\ldots,i_j\},\ i_j}\Big]$$

   Note that this quantity depends on M. With that, by Lemmas 3.4 and 3.5, given M, it depends only on the cardinalities $|L|$ and $|R \cap L|$.

3. Using the observation in the preceding paragraph, given $R \subseteq L \subseteq M \subseteq [k]$, with $|L| = \ell$, and $|R| = r$, we may also write $\mathcal{S}[\mu]^r_M\Big(\ell\Big)$ for $\mathcal{S}[\mu]^R_M(L)$.

   In particular, note that the proof of Lemma 3.5 gives, in this notation

$$\mathcal{S}[\mu]^r_{[k]}\Big(m\Big) \quad = \quad \sum_{j=1}^{m-r} y_j \quad + \quad \lambda \cdot \sum_{t=0}^{r-1} x^t_{m-r+t+1} \tag{15}$$

4. Finally, for $M \subseteq [k]$ and $0 \le r \le |M|$, we write

$$\mu^r_M \quad = \quad \mathop{\mathbb{E}}_{|R|=r, R\subseteq M} \quad \mu^R(M) \quad \text{and} \quad \mathcal{S}[\mu]^r_M \quad = \quad \mathop{\mathbb{E}}_{|R|=r, R\subseteq M} \quad \mathcal{S}[\mu]^R_M(M)$$

We have completed introducing the new notation. In this notation the claim of the theorem amounts to:

$$\mu^r_{[k]} \quad \le \quad \mathcal{S}[\mu]^r_{[k]} \tag{16}$$

We start with a lemma connecting the value of a solution of the optimization problem to these of smaller problems.

Lemma 3.6:

$$\mu^r_{[k]} \quad \le \quad \lambda \cdot \mu^{r-1}_{[k]} \quad + \quad (1-\lambda) \cdot \mathop{\mathbb{E}}_{i \in [k]} \; \mu^{r-1}_{[k]\setminus\{i\}} \tag{17}$$

Proof:

Since the feasible solution $\Big\{x^R_{S,i}\Big\}$ satisfies constraints 2 and 3 of (9), for any $i \in R \subseteq [k]$ holds $\mu^R\Big([k]\Big) \; = \; \mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big) \; + \; x^R_{[k],i}$.

Similarly, $\mu^{R\setminus\{i\}}\Big([k]\Big) \quad = \quad \mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big) \quad + \quad x^{R\setminus\{i\}}_{[k],i}$.

Hence, by constraint 4,

$$x^R_{[k],i} \;\le\; \lambda \cdot \left(x^{R\setminus\{i\}}_{[k],i}\right) \;=\; \lambda \cdot \left(\mu^{R\setminus\{i\}}\Big([k]\Big) \;-\; \mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big)\right)$$

Averaging,

$$\mu^r_{[k]} \quad = \quad \mathop{\mathbb{E}}_{R\subseteq[k],\ |R|=r} \quad \mu^R_{[k]} \quad = \quad \mathop{\mathbb{E}}_{R,\ i\in R} \quad \left(\mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big) \;+\; x^R_{[k],i}\right) \quad \le$$

$$\mathop{\mathbb{E}}_{R,\ i\in R} \quad \mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big) \quad + \quad \lambda \cdot \mathop{\mathbb{E}}_{R,\ i\in R} \quad \left(\mu^{R\setminus\{i\}}\Big([k]\Big) \;-\; \mu^{R\setminus\{i\}}\Big([k] \setminus \{i\}\Big)\right) \quad =$$

$$\lambda \cdot \mathop{\mathbb{E}}_{R,\ i \in R} \quad \mu^{R \setminus \{i\}}\Big([k]\Big) \quad + \quad (1-\lambda) \cdot \mathop{\mathbb{E}}_{R,\ i \in R} \quad \mu^{R \setminus \{i\}}\Big([k] \setminus \{i\}\Big)$$

It remains to note

$$\mathop{\mathbb{E}}_{R,\ i \in R} \quad \mu^{R \setminus \{i\}}\Big([k] \setminus \{i\}\Big) \quad = \quad \mathop{\mathbb{E}}_{i \in [k]} \quad \mathop{\mathbb{E}}_{|T|=r-1,\ T \subseteq [k] \setminus \{i\}} \quad \mu^{T}\Big([k] \setminus \{i\}\Big) \quad = \quad \mathop{\mathbb{E}}_{i \in [k]} \ \mu^{r-1}_{[k] \setminus \{i\}}$$

and, similarly, $\mathbb{E}_{R,\ i \in R} \ \mu^{R \setminus \{i\}}\Big([k]\Big) \ = \ \mu^{r-1}_{[k]}$.

▮

We now prove (16), starting from (17).

First, note that, by Lemma 3.4 and by the induction hypothesis for $k-1$, we have $\mu^{r-1}_{[k] \setminus \{i\}} \ \leq \ \mathcal{S}[\mu]^{r-1}_{[k] \setminus \{i\}}$, for all $i \in [k]$.

Next, note that, by the induction hypothesis for k and $r-1$, we have $\mu^{r-1}_{[k]} \ \leq \ \mathcal{S}[\mu]^{r-1}_{[k]}$.

This gives

$$\mu^{r}_{[k]} \quad \leq \quad \lambda \cdot \mathcal{S}[\mu]^{r-1}_{[k]} \quad + \quad (1-\lambda) \cdot \mathop{\mathbb{E}}_{i \in [k]} \ \mathcal{S}[\mu]^{r-1}_{[k] \setminus \{i\}}$$

This implies that to prove (16) it suffices to show the following two identities:

1. $$\mathop{\mathbb{E}}_{i \in [k]} \ \mathcal{S}[\mu]^{r-1}_{[k] \setminus \{i\}} \quad = \quad \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big)$$

2. $$\mathcal{S}[\mu]^{r}_{[k]} \quad = \quad \lambda \cdot \mathcal{S}[\mu]^{r-1}_{[k]} \quad + \quad (1-\lambda) \cdot \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big)$$

Lemma 3.7:

$$\mathop{\mathbb{E}}_{i \in [k]} \ \mathcal{S}[\mu]^{r-1}_{[k] \setminus \{i\}} \quad = \quad \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big)$$

Proof: We introduce the following notation. For $i = 1, ..., k$ and for $0 \leq r < s \leq k-1$, let

$$y_{s,i} \quad = \quad y_{s,\ [k] \setminus \{i\}} \quad \text{and} \quad x^{r}_{s,i} \quad = \quad x^{r}_{s,\ [k] \setminus \{i\}}$$

The values on the RHS of these identities are defined as in (12) and in (13) for the corresponding restricted problems.

We start with observing that $\mathbb{E}_{i \in [k]} \, y_{s,i} \ = \ y_s$. In fact, by definition,

$$\mathop{\mathbb{E}}_{i \in [k]} \ y_{s,i} \quad = \quad \mathop{\mathbb{E}}_{i \in [k]} \ \mathop{\mathbb{E}}_{|S|=s, S \subseteq [k] \setminus \{i\}, j \in S} \quad y_{S,j} \quad = \quad \mathop{\mathbb{E}}_{|S|=s, j \in S} \ y_{S,j} \quad = \quad y_s$$

Next, we claim that for all $0 \leq r < s \leq k-1$ holds $\mathbb{E}_{i \in [k]} \ x^{r}_{s,i} \ = \ x^{r}_{s}$.

This is easy to verify by induction on r. Note that we already know the claim holds for $r = 0$, and the induction step follows directly from the definitions and the induction hypothesis.

We now apply (14) to the restricted problems, to obtain that, for each $1 \le i \le k$ holds

$$\mathcal{S}[\mu]^{r-1}_{[k]\setminus\{i\}} \quad = \quad \sum_{j=1}^{k-r} y_{j,\ i} \quad + \quad \lambda \cdot \sum_{t=0}^{r-2} x^t_{k-r+t+1,\ i}$$

Hence, we have:

$$\mathop{\mathbb{E}}_{i\in[k]} \mathcal{S}[\mu]^{r-1}_{[k]\setminus\{i\}} \quad = \quad \sum_{j=1}^{k-r} \mathop{\mathbb{E}}_{i\in[k]} y_{j,\ i} \quad + \quad \lambda \cdot \sum_{t=0}^{r-2} \mathop{\mathbb{E}}_{i\in[k]} x^t_{k-r+t+1,\ i} \quad = \quad \sum_{j=1}^{k-r} y_j \ + \ \lambda \cdot \sum_{t=0}^{r-2} x^t_{k-r+t+1}$$

This, by (15), equals to $\mathcal{S}[\mu]^{r-1}_{[k]}(k-1)$, completing the proof of the lemma.

▮

Lemma 3.8:

$$\mathcal{S}[\mu]^r_{[k]} \quad = \quad \lambda \cdot \mathcal{S}[\mu]^{r-1}_{[k]} \quad + \quad (1-\lambda) \cdot \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big)$$

Proof:

The proof of this lemma is similar to that of Lemma 3.6.

Since the symmetric solution $\mathcal{S}_{[k]}\left[x^R_{S,i}\right]$ $\Big($ which is the same as $\left\{\bar{x}^R_{S,i}\right\}\Big)$ satisfies constraints 2 and 3 of (9), for any $i \in R \subseteq [k]$ holds

$$\mathcal{S}[\mu]^R_{[k]}\Big([k]\Big) \quad = \quad \mathcal{S}[\mu]^{R\setminus\{i\}}_{[k]}\Big([k]\setminus\{i\}\Big) \quad + \quad \mathcal{S}_{[k]}\left[x^R_{[k],i}\right]$$

Consider the notation we have introduced above. Using items 3 and 4 in the description of this notation, and recalling $\mathcal{S}_{[k]}\left[x^R_{[k],i}\right] = \lambda \cdot x^{r-1}_k$, we can rewrite this equality as

$$\mathcal{S}[\mu]^r_{[k]} \quad = \quad \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big) \quad + \quad \lambda \cdot x^{r-1}_k$$

On the other hand, we have, for $i \in R \subseteq [k]$:

$$\mathcal{S}[\mu]^{R\setminus\{i\}}_{[k]}\Big([k]\Big) \quad = \quad \mathcal{S}[\mu]^{R\setminus\{i\}}_{[k]}\Big([k]\setminus\{i\}\Big) \quad + \quad \mathcal{S}_{[k]}\left[x^{R\setminus\{i\}}_{[k],i}\right]$$

which is the same as

$$\mathcal{S}[\mu]^{r-1}_{[k]} \quad = \quad \mathcal{S}[\mu]^{r-1}_{[k]}\Big(k-1\Big) \quad + \quad x^{r-1}_k$$

Combining these two identities immediately implies the claim of the lemma.

▮

This completes the proof of (16) and of the theorem.

▮

3.3 The value of the symmetric optimization problem

Let $\left\{\bar{x}^R_{S,i}\right\}$ be the symmetric solution for the symmetric version of (9). By Corollary 3.3, its value depends linearly on the symmetric boundary data $y_1, ..., y_k$, since $\{x^r_t\}$ are fixed linear functions of $y_1, ..., y_k$. Let us denote this value by $V(y_1, ..., y_k)$.

For $1 \leq s \leq k$, let e_s be the initial data vector with $y_s = 1$ and all the remaining y_t vanishing. Then $V(y_1, \ldots, y_k) \;=\; \sum_{s=1}^k y_s \cdot V(e_s)$.

Next, we find the values of the parameters x^r_t for initial data given by a unit vector.

Lemma 3.9: *Let the initial data be given by the unit vector e_s, for some $1 \leq s \leq k$. Then the values of the parameters x^r_t, for $0 \leq r < t \leq k$, are as follows.*

$$x^r_t \quad = \quad \begin{cases} \binom{r}{t-s} \cdot \lambda^{r-(t-s)} \cdot (1-\lambda)^{t-s} & \textit{if} \qquad s \;\leq\; t \;\leq\; s+r \\ 0 & \textit{otherwise} \end{cases}$$

(We use the convention $\binom{0}{0} = 1$.)

Proof: The claim of the lemma is easily verifiable by induction on r, or by directly verifying that (13) holds. ▮

Corollary 3.10:

$$V(e_s) \quad = \quad \lambda^s \cdot \sum_{m=0}^{k-s} \binom{s+m-1}{m} \cdot (1-\lambda)^m \quad = \quad 1 \quad - \quad \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j}$$

Proof: The first equality follows from Corollary 3.3. For the second equality, we proceed as follows

$$V(e_s) \quad = \quad \frac{\lambda^s}{(s-1)!} \cdot \frac{\partial^{s-1}}{\partial x^{s-1}} \left[(1 + x + \ldots + x^{k-1}\right]_{x=1-\lambda} \quad =$$

$$\frac{\lambda^s}{(s-1)!} \cdot \left(\frac{\partial^{s-1}}{\partial x^{s-1}} \left[\frac{1}{1-x}\right]_{x=1-\lambda} \quad - \quad \frac{\partial^{s-1}}{\partial x^{s-1}} \left[\frac{x^k}{1-x}\right]_{x=1-\lambda} \right) \quad =$$

$$1 \quad - \quad \frac{\lambda^s}{(s-1)!} \cdot \frac{\partial^{s-1}}{\partial x^{s-1}} \left[\frac{x^k}{1-x}\right]_{x=1-\lambda}$$

We have

$$\frac{\partial^t}{\partial x^t} \left[\frac{x^k}{1-x}\right] \quad = \quad \sum_{i=0}^{t} \binom{t}{i} \frac{\partial^i}{\partial x^i} \left[\frac{1}{1-x}\right] \cdot \frac{\partial^{t-i}}{\partial x^{t-i}} \left[x^k\right] \quad =$$

$$\sum_{i=0}^{t} \binom{t}{i} \cdot i! \cdot \frac{k!}{(k-t+i)!} \cdot x^{k-t+i} \cdot \frac{1}{(1-x)^{i+1}}$$

Substituting $j = t - i$ and rearranging, this is

$$\frac{t!}{(1-x)^{t+1}} \cdot \sum_{j=0}^{t} \binom{k}{j} (1-x)^j \cdot x^{k-j}$$

Substituting $t = s - 1$, $x = 1 - \lambda$, and simplifying, we get

$$V(e_s) \quad = \quad 1 \quad - \quad \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j}$$

∎

Corollary 3.11:

$$V(y_1, \ldots, y_k) \quad = \quad \sum_{s=1}^{k} \left(1 \ - \ \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j} \right) \ \cdot \ y_s$$

4 Proof of Theorem 1.7

We start with introducing some more notation.

4.0.1 Notation

- For a subset S of $[n]$ of cardinality at most $n-2$, and for distinct $i, j \notin S$, we set

$$Z_{S;i,j} \quad = \quad Ent\Big(f \mid S \cup \{i,j\}\Big) \ - \ Ent\Big(f \mid S \cup \{i\}\Big) \ - \ Ent\Big(f \mid S \cup \{j\}\Big) \ + \ Ent\Big(f \mid S\Big)$$

- For $s = 1, ..., n-1$, let $t_s \ = \ \mathbb{E}_{S,i,j} \ Z_{S;i,j}$.

  Here the expectation is taken over all subsets S of $[n]$ of cardinality $s-1$, and, given S, over all distinct i, j not in S.

- Let A be a subset of $[n]$ of cardinality $k < n$ and let $m \notin A$. For $1 \leq s \leq k$, let

$$Y(A, m, s) \quad = \quad \mathop{\mathbb{E}}_{S,i} \ Z_{S;i,m}$$

  where the expectation goes over subsets $S \subseteq A$ of cardinality $s-1$, and over $i \in A \setminus S$.

- For $1 \leq s \leq k \leq n$ let

$$\Lambda(k,s,\lambda) \quad = \quad 1 \quad - \quad \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j}$$

Proposition 4.1: *Let f be a nonnegative function on $\{0,1\}^n$. Let A be a subset of $[n]$ of cardinality $k < n$ and let $m \notin A$.*

Then

$$I_{T_{e_A} f}\ (A,m) \quad \leq \quad \sum_{s=1}^{k} \ \Lambda(k,s,\lambda) \ \cdot \ Y(A,m,s)$$

Proof:

By Theorem 2.1, the value of $I_{T_{e_A} f}\ (A,m)$ is bounded by the value of the linear optimization problem (9), with appropriate changes of indices.

By Theorem 3.2, this last value is upperbounded by the value of the symmetric version of the problem, which, according to Corollary 3.11, and tracing out the appropriate changes in indices and notation, is given by $\sum_{s=1}^{k} \ \Lambda(k,s,\lambda) \cdot Y(A,m,s)$.

▮

Proof: (Of the theorem)

The proof relies on several lemmas. We start with a technical claim.

Lemma 4.2: *Let $1 \leq s \leq n-1$ be integer parameters. Let $0 < \lambda < 1$. Then*

$$\sum_{k=s}^{n-1} \Lambda(k,s,\lambda) \quad = \quad \left(n - \frac{s}{\lambda}\right) \ + \ \frac{1}{\lambda} \cdot \sum_{j=0}^{s-1} \ \sum_{t=0}^{j} \binom{n}{t} \lambda^t (1-\lambda)^{n-t}$$

Proof:

$$\sum_{k=s}^{n-1} \Lambda(k,s,\lambda) \quad = \quad \sum_{k=s}^{n-1} \left(1 \quad - \quad \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j}\right) \quad =$$

$$\Big(n-s\Big) \quad - \quad \sum_{k=s}^{n-1} \ \sum_{j=0}^{s-1} \binom{k}{j} \lambda^j (1-\lambda)^{k-j} \quad = \quad \Big(n-s\Big) \quad - \quad \sum_{j=0}^{s-1} \ \lambda^j \cdot \sum_{k=s}^{n-1} \ \binom{k}{j} (1-\lambda)^{k-j}$$

A simple calculation, similar to that in the proof of Corollary 3.10, gives

$$\lambda^j \cdot \sum_{k=s}^{n-1} \ \binom{k}{j} (1-\lambda)^{k-j} \quad = \quad \frac{1}{\lambda} \cdot \left(\sum_{t=0}^{j} \binom{s}{t} \lambda^t (1-\lambda)^{s-t} \quad - \quad \sum_{t=0}^{j} \binom{n}{t} \lambda^t (1-\lambda)^{n-t} \right)$$

The proof of the lemma is completed by summing the RHS over j, and observing

$$\sum_{j=0}^{s-1} \sum_{t=0}^{j} \binom{s}{t} \lambda^t (1-\lambda)^{s-t} \quad = \quad (1-\lambda) \cdot s$$

▮

Lemma 4.3: *Let f be a nonnegative function on $\{0,1\}^n$ with expectation 1. Then*

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad \sum_{i=1}^{n} \phi\Big(Ent\Big(f \mid \{i\}\Big)\Big) \quad + \quad \sum_{s=1}^{n-1} w_s \cdot t_s$$

where

$$w_s \quad = \quad \Big(\lambda n - s\Big) \; + \; \sum_{j=0}^{s-1} \sum_{t=0}^{j} \binom{n}{t} \lambda^t (1-\lambda)^{n-t}$$

Lemma 4.4: *Let f be a nonnegative function on $\{0,1\}^n$. For any $0 \le u \le n-1$ holds*

$$\mathop{\mathbb{E}}_{|B|=u+1} Ent\Big(f \mid B\Big) \; - \; (u+1) \cdot \mathop{\mathbb{E}}_{i \in [n]} Ent\Big(f \mid \{i\}\Big) \quad = \quad \sum_{s=1}^{u} \Big(u - s + 1\Big) \cdot t_s$$

Next, we derive the theorem, assuming Lemmas 4.3 and 4.4 to hold.

Let T be a random subset of $[n]$ generated by sampling each element $i \in [n]$ independently with probability λ. We will show

$$\mathop{\mathbb{E}}_{T} \left(Ent\Big(f \mid T\Big) \; - \; \sum_{i \in T} Ent\Big(f \mid \{i\}\Big) \right) \quad = \quad \sum_{s=1}^{n-1} w_s \cdot t_s$$

Combining this with the claim of Lemma 4.3 will complete the proof.

For $0 \le k \le n$, let $p_k = \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$. And, for $0 \le u \le n-1$, let

$$\mu_u \quad = \quad \mathop{\mathbb{E}}_{|B|=u+1} \left(Ent\Big(f \mid B\Big) \; - \; \sum_{i \in B} Ent\Big(f \mid \{i\}\Big) \right)$$

Then, using Lemma 4.4 and observing that $\mu_0 = 0$,

$$\mathop{\mathbb{E}}_{T} \left(Ent\Big(f \mid T\Big) - \sum_{i \in T} Ent\Big(f \mid \{i\}\Big) \right) \quad = \quad \sum_{k=2}^{n} p_k \; \mu_{k-1} \quad =$$

$$\sum_{k=2}^{n} p_k \sum_{s=1}^{k-1} (k-s)\cdot t_s \;=\; \sum_{s=1}^{n-1}\left(\sum_{k=s+1}^{n}(k-s)\,p_k\right)\cdot t_s$$

We conclude by verifying the identity $w_s = \sum_{k=s+1}^{n}(k-s)\,p_k$, for $s = 1, ..., n-1$.

In fact,

$$w_s \;=\; \Big(\lambda n - s\Big) \;+\; \sum_{j=0}^{s-1}\sum_{t=0}^{j} p_t \;=\; \sum_{k=0}^{n}(k-s)\,p_k \;+\; \sum_{t=0}^{s-1}(s-t)\,p_t \;=\; \sum_{k=s+1}^{n}(k-s)\,p_k$$

▮

It remains to prove the lemmas.

Proof: (Of Lemma 4.3)

Recall that, by the chain rule for noisy entropy (7), for any permutation $\sigma \in S_n$ holds that $Ent\Big(T_\epsilon f\Big)$ is bounded from above by

$$\sum_{i=1}^{n}\phi\left(Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}}f \mid \{\sigma(1),\ldots,\sigma(i)\}\Big) \;-\; Ent\Big(T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}}f \mid \{\sigma(1),\ldots,\sigma(i-1)\}\Big)\right)$$

Using the notation introduced in Subsection 1.4.1, we can write this as

$$\sum_{i=1}^{n}\phi\left(Ent\Big(f \mid \{\sigma(i)\}\Big) \;+\; I_{T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}}f}\left(\Big\{\sigma(1),\ldots,\sigma(i-1)\Big\},\ \sigma(i)\right)\right)$$

Observe that the function ϕ is concave, and $\phi(0) = 0$. Hence $\phi(x+y) \le \phi(x) + \phi(y)$ for any $0 \le x, y \le 1$. By this subbaditivity of ϕ, the last expression is at most

$$\sum_{i=1}^{n}\phi\left(Ent\Big(f \mid \{i\}\Big)\right) \;+\; \sum_{k=2}^{n}\phi\left(I_{T_{\epsilon_{\{\sigma(1),\ldots,\sigma(k-1)\}}}f}\left(\Big\{\sigma(1),\ldots,\sigma(k-1)\Big\},\ \sigma(k)\right)\right)$$

Averaging this expression over all $\sigma \in S_n$, we obtain

$$Ent\Big(T_\epsilon f\Big) \;\le\; \sum_{i=1}^{n}\phi\left(Ent\Big(f \mid \{i\}\Big)\right) \;+\; \mu,$$

where

$$\mu = \mathop{\mathbb{E}}_{\sigma}\ \sum_{k=2}^{n}\phi\left(I_{T_{\epsilon_{\{\sigma(1),\ldots,\sigma(k-1)\}}}f}\left(\Big\{\sigma(1),\ldots,\sigma(k-1)\Big\},\ \sigma(k)\right)\right)$$

Next, we upper bound μ. By transitivity of action of the symmetric group and by concavity of ϕ we have

$$\mu \quad \le \quad \sum_{k=1}^{n-1} \phi\Big(b_k\Big) \quad \text{where} \quad b_k \quad = \quad \mathop{\mathbb{E}}_{A,m} \; T_{\epsilon_A f} \, \Big(A, m\Big)$$

where the expectation is over all $A \subseteq [n]$ of cardinality k and $m \notin A$.

Applying Proposition 4.1, we get

$$b_k \quad \le \quad \mathop{\mathbb{E}}_{A,m} \; \sum_{s=1}^{k} \Lambda\Big(k, s, \lambda\Big) \cdot Y\Big(A, m, s\Big) \quad = \quad \sum_{s=1}^{k} \Lambda\Big(k, s, \lambda\Big) \cdot \mathop{\mathbb{E}}_{A,m} \; Y\Big(A, m, s\Big)$$

By the definition of $Y\Big(A, m, s\Big)$,

$$\mathop{\mathbb{E}}_{A,m} \; Y\Big(A, m, s\Big) \quad = \quad \mathop{\mathbb{E}}_{A,m} \; \mathop{\mathbb{E}}_{S,i} \; Z_{S;i,m} \quad = \quad \mathop{\mathbb{E}}_{S,i,m} \; Z_{S;i,m} \cdot \mathop{\mathbb{E}}_{A} \; 1 \quad = \quad \mathop{\mathbb{E}}_{S,i,m} \; Z_{S;i,m}$$

where in the second expression the first expectation is over k-subsets A of $[n]$ and $m \notin A$, and the second expectation is over $(s-1)$-subsets S of A and over $i \in A \setminus S$. Rearranging, we get the third expression in which the first expectation is over all subsets S of $[n]$ of cardinality $s-1$ and over all distinct $i, m \notin S$, and the second expectation is over all supersets A of S of cardinality k with $i \in A$ and $m \notin A$.

Recalling the definition of t_s above, we deduce $b_k \quad = \quad \sum_{s=1}^{k} \Lambda\Big(k, s, \lambda\Big) \cdot t_s$.

Using the inequality $\phi(x) \le \lambda x$, and Lemma 4.2, we have

$$\mu \quad \le \quad \lambda \cdot \sum_{k=1}^{n-1} b_k \quad = \quad \lambda \cdot \sum_{k=1}^{n-1} \; \sum_{s=1}^{k} \Lambda\Big(k, s, \lambda\Big) \cdot t_s \quad =$$

$$\sum_{s=1}^{n-1} t_s \cdot \left(\lambda \cdot \sum_{k=s}^{n-1} \Lambda\Big(k, s, \lambda\Big)\right) \quad = \quad \sum_{s=1}^{n-1} w_s \cdot t_s$$

∎

Proof: (of Lemma 4.4)

By (10), for any subset A of $[n]$ of cardinality $1 \le k \le n-1$, for any $m \notin A$, and for any bijection $\tau : [k] \rightarrow A$ holds, in the notation of this section,

$$\sum_{s=1}^{k} \; Z_{\{\tau(1),\ldots,\tau(s-1)\};\tau(s),m} \quad = \quad I_f\Big(A, m\Big)$$

We now average over all the variables, setting

$$c_k \quad = \quad \mathop{\mathbb{E}}_{A,m,\tau} \; \sum_{s=1}^{k} Z_{\{\tau(1),\ldots,\tau(s-1)\};\tau(s),m}$$

On one hand, we have

$$c_k \quad = \quad \mathop{\mathbb{E}}_{A,m} I_f\Big(A,m\Big) \quad = \quad \mathop{\mathbb{E}}_{A,m} \Bigg(Ent\Big(f \mid A \cup \{m\}\Big) - Ent\Big(f \mid A\Big) - Ent\Big(f \mid \{m\}\Big)\Bigg) \quad =$$

$$\mathop{\mathbb{E}}_{|B|=k+1} Ent\Big(f \mid B\Big) \; - \; \mathop{\mathbb{E}}_{|A|=k} Ent\Big(f \mid A\Big) \; - \; \mathop{\mathbb{E}}_{i\in[n]} Ent\Big(f \mid \{i\}\Big)$$

On the other hand, similarly to the computation in the preceding lemma, we have

$$c_k \quad = \quad \sum_{s=1}^{k} \mathop{\mathbb{E}}_{A,m,\tau} Z_{\{\tau(1),\ldots,\tau(s-1)\};\tau(s),m} \quad = \quad \sum_{s=1}^{k} \mathop{\mathbb{E}}_{A,S,i,m} Z_{S;i,m} \quad = \quad \sum_{s=1}^{k} t_s$$

where the expectation in the third expression is over k-subsets A of $[n]$, over $(s-1)$-subsets S of A, over $m \notin A$ and $i \in A \setminus S$.

Hence, for any $1 \leq u \leq n-1$ holds

$$\mathop{\mathbb{E}}_{|B|=u+1} Ent\Big(f \mid B\Big) \; - \; (u+1) \cdot \mathop{\mathbb{E}}_{i\in[n]} Ent\Big(f \mid \{i\}\Big) \quad = \quad \sum_{k=1}^{u} c_k \quad = \quad \sum_{s=1}^{u} \Big(u-s+1\Big) \cdot t_s$$

completing the proof of the lemma and of the theorem.

∎

5 Proof of Theorem 1.14

Let δ be the constant in the theorem. We will assume in the following argument that δ is sufficiently small.

Let $0 < \epsilon < 1/2$ be a noise parameter, such that $(1-2\epsilon)^2 \leq \delta$. Let $\lambda = (1-2\epsilon)^2$.

Let $f : \{0,1\}^n \rightarrow \{0,1\}$ be a boolean function, satisfying the constraints of the theorem. Let $1 \leq k \leq n$ be the coordinate such that $|\widehat{f}(k)|$ is large. W.l.o.g. assume that $k = 1$ and that $\widehat{f}(1)$ is positive.

We introduce some additional notation.

Notation

- Let $0 \le \alpha \le \delta$ be such that $\widehat{f}(1) = (1-\alpha) \cdot \mathbb{E} f$.
- Let $0 \le \beta \le \delta$ be such that $\mathbb{E} f = 1/2 - \beta$. Let $\gamma = \alpha + \beta$.
- If $\alpha \le \lambda$, we define $\tau = \left(\frac{1-\lambda}{1-\alpha}\right)^2$, and define auxiliary noise ϵ_τ, such that $\Big(1 - 2\epsilon_\tau\Big)^2 = \tau$. If $\alpha > \lambda$, we set $\tau = 1$ and $\epsilon_\tau = 0$.
- Let ϵ_1 be such that $T_\epsilon = T_{\epsilon_1} T_{\epsilon_\tau}$. Let $\lambda_1 = (1-2\epsilon_1)^2$. Note that $\lambda = \tau \cdot \lambda_1$.
- Let $h = T_{\epsilon_\tau} f$. Note that $T_\epsilon f = T_{\epsilon_1} h$, and hence $Ent\Big(T_\epsilon f\Big) = Ent\Big(T_{\epsilon_1} h\Big)$.

5.1 Proof of the first claim of the theorem

We start with applying Theorem 1.7 to the function h with noise ϵ_1. The theorem is stated for functions with expectation 1. We modify it, using the linearity of entropy, to obtain

$$Ent\Big(T_{\epsilon_1} h\Big) \;\le\; \mathop{\mathbb{E}}_{T} \left(Ent\Big(h \mid T\Big) - \sum_{i \in T} Ent\Big(h \mid \{i\}\Big)\right) \;+\; \mathbb{E}\, h \cdot \sum_{i=1}^{n} \phi\left(Ent\left(\frac{h}{\mathbb{E}\, h} \;\Big|\; \{i\}\right),\; \epsilon_1\right)$$

Here T is a random subset of $[n]$ generated by sampling each element $i \in [n]$ independently with probability $(1-2\epsilon_1)^2$.

Since there are more than one noise parameters involved, we now write the function ϕ with the noise parameter stated explicitly.

Next, note that by (3), for any $1 \le i \le n$ holds

$$\mathbb{E}\, h \cdot \phi\left(Ent\left(\frac{h}{\mathbb{E}\, h} \;\Big|\; \{i\}\right),\; \epsilon_1\right) \;\le\; \lambda_1 \cdot Ent\Big(h \mid \{i\}\Big)$$

Hence the previous inequality implies

$$Ent\Big(T_{\epsilon_1} h\Big) \;\le\; \lambda_1 \cdot \mathop{\mathbb{E}}_{T, 1 \in T} \left(Ent\Big(h \mid T\Big) - Ent\Big(h \mid \{1\}\Big)\right) \;+$$

$$(1-\lambda_1) \cdot \mathop{\mathbb{E}}_{T, 1 \notin T} Ent\Big(h \mid T\Big) \;+\; \mathbb{E}\, h \cdot \phi\left(Ent\left(\frac{h}{\mathbb{E}\, h} \;\Big|\; \{1\}\right),\; \epsilon_1\right) \tag{18}$$

The proof will be based on three lemmas, which upperbound each of the three summands on the RHS of (18).

Lemma 5.1:

$$\mathop{\mathbb{E}}_{T,1\in T} \left(Ent\Big(h \mid T\Big) - Ent\Big(h \mid \{1\}\Big)\right) \quad \le \quad O\left(\lambda_1 \cdot \gamma \;+\; \gamma^2 \ln\left(\frac{1}{\gamma}\right)\right)$$

Lemma 5.2:

$$\mathop{\mathbb{E}}_{T,1\notin T} Ent\Big(h \mid T\Big) \quad \le \quad O\left(\lambda_1^2 \cdot \gamma \;+\; \lambda_1 \cdot \gamma^2 \ln\left(\frac{1}{\gamma}\right)\right)$$

Lemma 5.3:

$$\mathbb{E}\, h \cdot \phi\left(Ent\left(\frac{h}{\mathbb{E}\, h} \,\Big|\, \{1\}\right),\; \epsilon_1\right) \quad \le \quad \frac{1}{2} \cdot \Big(1 - H(\epsilon)\Big) \;-\; \Omega\Big(\lambda \cdot \gamma\Big)$$

The asymptotic notation in each of the lemmas hides absolute constants.

Given the lemmas, the first claim of the theorem is easy to verify. Indeed, recall that λ_1 is a constant multiple of λ. Hence, the lemmas and (18) imply that

$$Ent\Big(T_\epsilon f\Big) \quad = \quad Ent\Big(T_{\epsilon_1} h\Big) \quad \le \quad \frac{1}{2} \cdot \Big(1 - H(\epsilon)\Big) \;-\; \Omega\Big(\lambda \cdot \gamma\Big) \;+\; o_{\lambda,\gamma \rightarrow 0}\Big(\lambda \cdot \gamma\Big)$$

Therefore, for a sufficiently small $\delta > 0$, bearing in mind that $0 \le \alpha, \beta, \lambda \le \delta$, the claim holds.

It remains to prove the lemmas. For that purpose we will need the following version of the logarithmic Sobolev inequality for the boolean cube.

Lemma 5.4: *Let g be a nonnegative function on $\{0,1\}^n$. Let $\mathcal{E}(g,g)$ be the Dirichlet form, given by $\mathcal{E}(g,g) = \mathbb{E}_{x \in \{0,1\}^n} \mathbb{E}_{y \sim x} \Big(g(y) - g(x)\Big)^2$. Then*

$$\mathcal{E}(g,g) \quad \ge \quad 2 \ln 2 \cdot \mathbb{E}\, g \cdot Ent\Big(g\Big)$$

Proof:

We start with a simple auxiliary claim.

Let $x_1 \ge x_2 \ge ... \ge x_N$ be nonnegative numbers summing to 1. Then the numbers $y_k = \frac{x_k^2}{\sum_{i=1}^N x_i^2}$, for $k = 1, ..., N$, majorize $\{x_k\}$. That is,

$$y_1 \;\ge\; x_1, \quad y_1 + y_2 \;\ge\; x_1 + x_2, \quad \ldots \quad , y_1 + ... + y_N \;=\; 1 \;=\; x_1 + ... + x_N$$

To see this, fix some $1 \le t \le N$. We have to show $\sum_{k=1}^t x_k^2 \;\ge\; \left(\sum_{k=1}^t x_k\right) \cdot \left(\sum_{k=1}^N x_k^2\right)$.

We may and will assume that all of the x_k are strictly positive. After some rearrangement, the claim reduces to showing

$$\frac{\sum_{k=1}^{t} x_k^2}{\sum_{k=1}^{t} x_k} \quad \geq \quad \frac{\sum_{m=t+1}^{N} x_m^2}{\sum_{k=t+1}^{N} x_m}$$

This holds because the LHS is lowerbounded by x_t, and the RHS is upperbounded by x_{t+1}.

A simple corollary of this claim is that for any nonnegative not identically zero function g on a finite domain endowed with uniform measure, holds that $g^2/\mathbb{E}\, g^2$ majorizes $g/\mathbb{E}\, g$.

This is well-known to imply (see [10]) that $g/\mathbb{E}\, g$ is a convex combination of permuted versions of $g^2/\mathbb{E}\, g^2$. Since the entropy functional is linear and convex, this implies

$$Ent\Big(g^2\Big) \quad \geq \quad \frac{\mathbb{E}\, g^2}{\mathbb{E}\, g} \cdot Ent\Big(g\Big) \quad \geq \quad \mathbb{E}\, g \cdot Ent\Big(g\Big)$$

The claim of the lemma follows from this inequality combined with the logarithmic Sobolev inequality [8]:

$$\mathcal{E}(g,g) \quad \geq \quad 2\ln 2 \cdot Ent\Big(g^2\Big)$$

∎

In the following argument we are going to use the Walsh-Fourier expansion for functions on the boolean cube, writing a function g as $\sum_{S\subseteq[n]} \widehat{g}(S) \cdot W_S$, where $\Big\{W_S\Big\}_{S\subseteq[n]}$ is the Walsh-Fourier basis.

In particular, for the Dirichlet form, we have $\mathcal{E}(g,g) = 4 \cdot \sum_{S\subseteq[n]} |S| \widehat{g}^2(S)$. Hence the preceding lemma implies

$$Ent\Big(g\Big) \quad \leq \quad \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E}\, g} \cdot \sum_{S\subseteq[n]} |S|\; \widehat{g}^2(S) \tag{19}$$

We will also need the following precise version of an inequality of [4], due to [6]:

Theorem 5.5: *There exists a universal constant* $L > 0$ *with the following property. For* $g:\ \{0,1\}^n \to \{-1,1\}$, *let* $\rho = \Big(\sum_{A\subseteq[n]:|A|\geq 2} \widehat{g}^2(A)\Big)^{1/2}$. *Then there exists some* $B \subseteq [n]$ *with* $|B| \leq 1$ *such that*

$$\sum_{A\subseteq[n]:|A|\leq 1, A\neq B} \widehat{g}^2(A) \quad \leq \quad L \cdot \rho^4 \ln\left(\frac{2}{\rho}\right)$$

and $|\widehat{g}(B)|^2 \geq 1 \ - \ \rho^2 \ - \ L \cdot \rho^4 \ln\left(\frac{2}{\rho}\right)$.

Consider the function f and recall that it satisfies the assumptions of Theorem 1.14.

Let $g = 2f - 1$. Then $g : \{0,1\}^n \to \{-1,1\}$. Note that $\widehat{g}(0) = 2\widehat{f}(0) - 1$, and that $\widehat{g}(S) = 2\widehat{f}(S)$, for $|S| > 0$.

In particular, $\widehat{g}(0) = 2\,\mathbb{E}\,f - 1 = -2\beta$, and $\widehat{g}(\{1\}) = 2(1-\alpha)\,\mathbb{E}\,f = (1-\alpha)(1-2\beta)$.

Recall that $0 \le \alpha, \beta \le \delta$, and that $\gamma = \alpha + \beta$. Hence, assuming δ is sufficiently small, we have

$$\sum_{|A|\ge 2} \widehat{f}^2(A) \;\le\; \sum_{|A|\ge 2} \widehat{g}^2(A) \;\le\; 1 \;-\; \widehat{g}^2(\{1\}) \;\le\; L \cdot \gamma, \tag{20}$$

for some absolute constant L.

Applying Theorem 5.5 to the function g, we get, for a sufficiently large constant L_1,

$$\sum_{k=2}^{n} \widehat{f}^2\Big(\{k\}\Big) \;\le\; \sum_{k=2}^{n} \widehat{g}^2\Big(\{k\}\Big) \;\le\; L_1 \cdot \gamma^2 \ln\left(\frac{1}{\gamma}\right) \tag{21}$$

Proof of Lemma 5.2

Fix $T \subseteq [n]$. Let $g_T = \mathbb{E}\Big(h \mid T\Big)$.

Note that $g_T = \sum_{S \subseteq T} \widehat{h}(S) \cdot W_S$, and hence, by (19), we have

$$Ent\Big(g_T\Big) \;\le\; \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E}\, g_T} \cdot \sum_{S \subseteq T} |S|\; \widehat{h}^2(S) \;=\; \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E}\, h} \cdot \sum_{S \subseteq T} |S|\; \widehat{h}^2(S)$$

Hence,

$$\mathop{\mathbb{E}}_{T, 1 \notin T}\; Ent\Big(g_T\Big) \;\le\; \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E}\, h} \cdot \mathop{\mathbb{E}}_{T, 1 \notin T} \sum_{S \subseteq T} |S|\; \widehat{h}^2(S) \;=\; \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E}\, h} \cdot \sum_{S, 1 \notin S} |S| \lambda_1^{|S|}\; \widehat{h}^2(S)$$

Recall that $h = T_{\epsilon_\tau} f$. This means (see, e.g., [11]) that for any $S \subseteq [n]$, holds $\widehat{h}(S) = \tau^{|S|/2} \cdot \widehat{f}(S)$. In particular, $|\widehat{h}(S)| \le |\widehat{f}(S)|$. Applying (20) and (21), we have that, for a sufficiently large absolute constant L, the last expression is bounded by

$$L \cdot \left(\lambda_1^2 \cdot \gamma \;+\; \lambda_1 \cdot \gamma^2 \ln\left(\frac{1}{\gamma}\right) \right)$$

This concludes the proof of the lemma. ∎

Proof of Lemma 5.3

Let $g = \mathbb{E}\left(\frac{f}{\mathbb{E}f} \;\middle|\; \{1\}\right)$. Then g is a function on a 2-point space $\{0,1\}$, with $g(0) = 2-\alpha$ and $g(1) = \alpha$.

Observe that the noise operator commutes with the projection operator. Hence, since $h = T_{\epsilon_\tau} f$, we have $g_1 := \mathbb{E}\left(\frac{h}{\mathbb{E}h} \;\middle|\; \{1\}\right) = T_{\epsilon_\tau} g$.

Observe also that, by the definition of Mrs. Gerber's function ϕ, we have

$$\phi\Bigg(Ent\left(\frac{h}{\mathbb{E}\,h} \;\middle|\; \{1\}\right),\ \epsilon_1\Bigg) \quad = \quad Ent\Big(T_{\epsilon_1} g_1\Big) \quad = \quad Ent\Big(T_{\epsilon_1} T_{\epsilon_\tau} g\Big) \quad = \quad Ent\Big(T_{\epsilon} g\Big)$$

The last equality follows from the definition of ϵ_1 and ϵ_τ.

It is easy to verify that $T_\epsilon g(0) = 1 + (1-\alpha)\cdot\lambda^{1/2}$ and that $T_\epsilon g(1) = 1 - (1-\alpha)\cdot\lambda^{1/2}$.

Hence, $Ent\Big(T_\epsilon g\Big) = 1 - H_2\left(\frac{1-(1-\alpha)\cdot\lambda^{1/2}}{2}\right)$.

Recall that

$$H_2\left(\frac{1-x}{2}\right) \quad = \quad 1 \quad - \quad \frac{1}{\ln 2}\cdot\sum_{k=1}^{\infty}\frac{1}{2k(2k-1)}\cdot x^{2k}$$

with the series converging absolutely for $-1 \le x \le 1$.

Let $F(x) = 1 - H_2\left(\frac{1-\sqrt{x}}{2}\right)$, for $0 \le x \le 1$. Then $F(x) = \frac{1}{\ln 2}\cdot\sum_{k=1}^{\infty}\frac{1}{2k(2k-1)}\cdot x^k$.

This is a convex function on $[0,1]$, and hence for any $0 \le x < y \le 1$ holds $F(y) - F(x) \ge (y-x)\cdot F'(x)$. The derivative F' is given by $F'(x) = \frac{1}{2\ln 2}\cdot\sum_{k=1}^{\infty}\frac{1}{2k-1}\cdot x^{k-1}$, with the series converging for $0 \le x < 1$.

Hence $F' \ge \frac{1}{2\ln 2}$ on $(0,1)$, and $F(y) - F(x) \ge \frac{1}{2\ln 2}\cdot(y-x)$. Applying this with $y = \lambda$ and $x = (1-\alpha)^2\cdot\lambda$, we get

$$\Big(1 - H_2(\epsilon)\Big) \;-\; Ent\Big(T_\epsilon g\Big) \quad = \quad F(\lambda) \;-\; F\Big((1-\alpha)^2\cdot\lambda\Big) \quad \ge \quad c_1\cdot\lambda\cdot\alpha$$

where $c_1 > 0$ is an absolute constant.

In other words,

$$\phi\Bigg(Ent\left(\frac{h}{\mathbb{E}\,h} \;\middle|\; \{1\}\right),\ \epsilon_1\Bigg) \quad = \quad Ent\Big(T_\epsilon g\Big) \quad \le \quad \Big(1 - H_2(\epsilon)\Big) \;-\; c_1\cdot\lambda\cdot\alpha$$

To conclude the proof of the lemma, note that, for a sufficiently small λ, we have $Ent\Big(T_\epsilon g\Big) \ge c_2\cdot\lambda$, for an absolute constant c_2, and hence

$$\mathbb{E}\,h\cdot\phi\Bigg(Ent\left(\frac{h}{\mathbb{E}\,h} \;\middle|\; \{1\}\right),\ \epsilon_1\Bigg) \quad = \quad \left(\frac{1}{2}-\beta\right)\cdot Ent\Big(T_\epsilon g\Big) \quad \le$$

$$\frac{1}{2} \cdot \Big(1 - H_2(\epsilon)\Big) \; - \; c \cdot \lambda \cdot (\alpha + \beta) \;\; = \;\; \frac{1}{2} \cdot \Big(1 - H_2(\epsilon)\Big) \; - \; c \cdot \lambda \cdot \gamma$$

for an absolute constant c. For the inequality, note that $1 - H_2(\epsilon) = F(\lambda) \geq \frac{1}{2 \ln 2} \cdot \lambda$.
This completes the proof of the lemma. ▮

The proof of Lemma 5.1 is somewhat harder. We present it in the next subsection.

5.1.1 Proof of Lemma 5.1

We proceed similarly to the proof of Lemma 5.2, and use the notation introduced in that proof. Given a function g on the boolean cube, we write $\mathbb{E}\Big(g \mid x_1 = 0, x_2, ..., x_k\Big)$ for the restriction of $\mathbb{E}\Big(g \mid x_1, x_2, ..., x_k\Big)$ on the subcube $x_1 = 0$, and similarly for $\mathbb{E}\Big(g \mid x_1 = 1, x_2, ..., x_k\Big)$.

We note that for $g = \sum_{S \subseteq [n]} \widehat{g}(S) \cdot W_S$, we have

$$\mathbb{E}\Big(g \mid x_1 = 0, x_2, ..., x_n\Big) \;\; = \;\; \sum_{R \subseteq [n], 1 \notin R} \Big(\widehat{g}(R) \; + \; \widehat{g}(R \cup \{1\})\Big) \cdot W_R$$

and

$$\mathbb{E}\Big(g \mid x_1 = 1, x_2, ..., x_n\Big) \;\; = \;\; \sum_{R \subseteq [n], 1 \notin R} \Big(\widehat{g}(R) \; - \; \widehat{g}(R \cup \{1\})\Big) \cdot W_R$$

We will also use the following easily verifiable identity, holding for nonnegative functions g:

$$Ent\Big(g\Big) \; - \; Ent\Big(g \mid \{1\} \Big) \;\; = \;\; \frac{1}{2} \cdot Ent\Big(g \mid x_1 = 0, x_2, ..., x_n\Big) \; + \; \frac{1}{2} \cdot Ent\Big(g \mid x_1 = 1, x_2, ..., x_n\Big)$$

As before, let $g_T = \mathbb{E}\Big(h \mid T\Big)$, for a subset $T \subseteq [n]$. Note that if $1 \in T$, then $\mathbb{E}\Big(g_T \mid \{1\}\Big) = \mathbb{E}\Big(h \mid \{1\}\Big)$.

Hence

$$\mathop{\mathbb{E}}_{T, 1 \in T} \Big(Ent\Big(h \mid T\Big) - Ent\Big(h \mid \{1\}\Big)\Big) \;\; = \;\; \mathop{\mathbb{E}}_{T, 1 \in T} \Big(Ent\Big(g_T\Big) - Ent\Big(g_T \mid \{1\}\Big)\Big) \;\; =$$

$$\frac{1}{2} \cdot \mathop{\mathbb{E}}_{T, 1 \in T} Ent\Big(g_T \mid x_1 = 0, x_2, ..., x_n\Big) \; + \; \frac{1}{2} \cdot \mathop{\mathbb{E}}_{T, 1 \in T} Ent\Big(g_T \mid x_1 = 1, x_2, ..., x_n\Big)$$

We will prove the lemma by showing that, for a sufficiently large absolute constant L, hold both

$$\mathop{\mathbb{E}}_{T, 1 \in T} Ent\Big(g_T \mid x_1 = 0, x_2, ..., x_n\Big) \;\; \leq \;\; L \cdot \lambda_1 \cdot \gamma \tag{22}$$

and

$$\mathop{\mathbb{E}}_{T, 1 \in T} Ent\Big(g_T \mid x_1 = 1, x_2, ..., x_n\Big) \;\; \leq \;\; L \cdot \left(\lambda_1 \cdot \gamma \; + \; \gamma^2 \ln\left(\frac{1}{\gamma}\right)\right) \tag{23}$$

Proof of (22)

Fix a subset $T \subseteq [n]$, with $1 \in T$. Recall that $g_T = \sum_{S \subseteq T} \widehat{h}(S) \cdot W_S$, and hence

$$\mathbb{E}\Big(g_T \mid x_1 = 0, x_2, ..., x_n\Big) = \sum_{R \subseteq T \setminus \{1\}} \Big(\widehat{h}(R) \ + \ \widehat{h}(R \cup \{1\})\Big) \cdot W_R$$

In particular,

$$\mathbb{E}\Big(g_T \mid x_1 = 0\Big) \quad = \quad \widehat{h}(0) \ + \ \widehat{h}(\{1\}) \quad = \quad \widehat{f}(0) \ + \ \tau^{1/2} \cdot \widehat{f}(\{1\}) \quad \geq \quad \mathbb{E} f$$

Applying (19), we have, for a sufficiently large constant L_1,

$$Ent\Big(g_T \mid x_1 = 0, x_2, ..., x_n\Big) \quad \leq \quad \frac{2}{\ln 2} \cdot \frac{1}{\mathbb{E} f} \cdot \sum_{R \subseteq T \setminus \{1\}} |R| \cdot \Big(\widehat{h}(R) \ + \ \widehat{h}(R \cup \{1\})\Big)^2 \quad \leq$$

$$L_1 \cdot \sum_{R \subseteq T \setminus \{1\}} |R| \cdot \Big(\widehat{h}^2(R) \ + \ \widehat{h}^2(R \cup \{1\})\Big)$$

Averaging over T, we have

$$\mathop{\mathbb{E}}_{T, 1 \in T} Ent\Big(g_T \mid x_1 = 0, x_2, ..., x_n\Big) \ \leq \ L_1 \cdot \left(\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{h}^2(R) \ + \ \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{h}^2(R \cup \{1\})\right)$$

Using the fact that $|\widehat{h}(S)| \leq |\widehat{f}(S)|$ for all $S \subseteq [n]$, and applying (20) and (21), we have, for a sufficiently large constant L_2,

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{h}^2(R) \quad \leq \quad L_2 \cdot \left(\lambda_1 \cdot \gamma^2 \ln\left(\frac{1}{\gamma}\right) \ + \ \lambda_1^2 \cdot \gamma\right)$$

and

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{h}^2(R \cup \{1\}) \quad \leq \quad L_2 \cdot \lambda_1 \cdot \gamma$$

Summing up, this gives (22).

Proof of (23)

Similarly to the above,

$$\mathbb{E}\left(g_T \mid x_1 = 1, x_2, ..., x_n\right) = \sum_{R \subseteq T \setminus \{1\}} \left(\widehat{h}(R) \; - \; \widehat{h}(R \cup \{1\})\right) \cdot W_R$$

Which means that

$$\mathbb{E}\left(g_T \mid x_1 = 1\right) \;\; = \;\; \widehat{h}(0) \; - \; \widehat{h}(\{1\}) \;\; = \;\; \widehat{f}(0) \; - \; \tau^{1/2} \cdot \widehat{f}(\{1\}) \;\; = \;\; \mathbb{E}\, f \cdot \left(1 \; - \; \tau^{1/2} \cdot (1 - \alpha)\right)$$

Recall that $\tau^{1/2} = 1$ if $\alpha \geq \lambda$ and $\tau^{1/2} = \frac{1-\lambda}{1-\alpha}$ otherwise. In both cases, note that we have $\mathbb{E}\left(g_T \mid x_1 = 1\right) \geq \lambda \cdot \mathbb{E}\, f$.

Applying (19), and averaging over T, we have, for a sufficiently large constant L_1,

$$\mathbb{E}_{T, 1 \in T} \; Ent\Big(g_T \mid x_1 = 1, x_2, ..., x_n\Big) \;\; \leq \;\; L_1 \cdot \frac{1}{\lambda} \cdot \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \cdot \left(\widehat{h}(R) \; - \; \widehat{h}(R \cup \{1\})\right)^2$$

Let $g = \mathbb{E}\left(h \mid x_1 = 1, x_2, ..., x_n\right)$. Then $g = \sum_{R \subseteq [n], 1 \notin R} \left(\widehat{h}(R) \; - \; \widehat{h}(R \cup \{1\})\right) \cdot W_R$. Hence

$$\mathbb{E}_{T, 1 \in T} \; Ent\Big(g_T \mid x_1 = 1, x_2, ..., x_n\Big) \;\; \leq \;\; L_1 \cdot \frac{1}{\lambda} \cdot \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \cdot \widehat{g}^2(R) \tag{24}$$

Consider the function g. Since $h = T_{\epsilon_\tau} f$, we have

$$g \;\; = \;\; \epsilon_\tau \cdot T_{\epsilon_\tau}\left(\mathbb{E}\left(f \mid x_1 = 0, x_2, ..., x_n\right)\right) \; + \; \left(1 \; - \; \epsilon_\tau\right) \cdot T_{\epsilon_\tau}\left(\mathbb{E}\left(f \mid x_1 = 1, x_2, ..., x_n\right)\right)$$

For $i = 0, 1$, let $f_i = \mathbb{E}\left(f \mid x_1 = i, x_2, ..., x_n\right)$, and let $t_i = T_{\epsilon_\tau} f_i$. Note that for $i = 0, 1$ and for any R, $1 \notin R$, holds $|\widehat{t_i}(R)| \;\leq\; |\widehat{f_i}(R)|$.

Therefore, since $g = \epsilon_\tau \cdot t_0 + \left(1 - \epsilon_\tau\right) \cdot t_1$, we have, for any R, $1 \notin R$ that

$$\widehat{g}^2(R) \;\; \leq \;\; \epsilon_\tau \cdot \widehat{t_0}^2(R) \; + \; \left(1 \; - \; \epsilon_\tau\right) \cdot \widehat{t_1}^2(R) \;\; \leq \;\; \epsilon_\tau \cdot \widehat{f_0}^2(R) \; + \; \left(1 \; - \; \epsilon_\tau\right) \cdot \widehat{f_1}^2(R)$$

Hence,

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \cdot \widehat{g}^2(R) \;\; \leq \;\; \epsilon_\tau \cdot \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{f_0}^2(R) \; + \; \left(1 - \epsilon_\tau\right) \cdot \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{f_1}^2(R) \tag{25}$$

Exactly as above, we have the following upper bound for the first summand: For a sufficiently large constant L_2 holds

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{f_0}^2(R) \ = \ \sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \Big(\widehat{f}(R) \ + \ \widehat{f}(R \cup \{1\}) \Big)^2 \ \leq \ L_2 \cdot \lambda_1 \cdot \gamma$$

Consider the second summand. The function f_1 is a boolean function, whose expectation equals $\widehat{f}(0) - \widehat{f}(\{1\}) = \alpha \cdot \mathbb{E} f \leq \alpha$. Similarly, $\mathbb{E} f_1^2 = \mathbb{E} f_1 \leq \alpha$.

We now apply the inequality of [20], which states that

For a boolean function $g : \ \{0,1\}^m \rightarrow \{0,1\}$ with expectation $\mu \leq 1/2$ holds $\sum_{k=1}^m \widehat{g}^2(\{k\}) \leq L_3 \cdot \mu^2 \cdot \ln(1/\mu)$, for a sufficiently large absolute constant L_3.

In our case, this implies $\sum_{k=2}^n \widehat{f_1}^2 \Big(\{k\} \Big) \leq L_3 \cdot \alpha^2 \cdot \ln\left(\frac{1}{\alpha}\right)$, for a sufficiently large constant L_3.

This means that, for a sufficiently large constant L_4, we can upperbound the second summand in (25) by

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \widehat{f_1}^2(R) \ \leq \ L_4 \cdot \left(\lambda_1 \cdot \alpha^2 \ln\left(\frac{1}{\alpha}\right) \ + \ \lambda_1^2 \cdot \alpha \right)$$

Recall that for $\alpha < \lambda$, we have $\epsilon_\tau = \frac{1 - \tau^{1/2}}{2} = \frac{1 - (1-\lambda)/(1-\alpha)}{2} \leq L_5 \cdot \lambda$, for an absolute constant L_5; and that for $\alpha \geq \lambda$, we have $\epsilon_\tau = 0$. Plugging these estimates into (25), we have

$$\sum_{R, 1 \notin R} |R| \lambda_1^{|R|} \cdot \widehat{g}^2(R) \ \leq \ L_2 \cdot L_5 \cdot \lambda \cdot \lambda_1 \cdot \gamma \ + \ L_4 \cdot \left(\lambda_1 \cdot \alpha^2 \ln\left(\frac{1}{\alpha}\right) \ + \ \lambda_1^2 \cdot \alpha \right)$$

And hence, coming back to (24), and recalling that $\lambda = \tau \cdot \lambda_1$, we have, for sufficiently large absolute constants L, L', that

$$\mathop{\mathbb{E}}_{T, 1 \in T} \ Ent\Big(g_T \mid x_1 = 1, x_2, ..., x_n \Big) \ \leq \ L' \cdot \left(\lambda_1 \cdot \gamma \ + \ \alpha^2 \ln\left(\frac{1}{\alpha}\right) \ + \ \lambda_1 \cdot \alpha \right) \ \leq$$

$$L \cdot \left(\lambda_1 \cdot \gamma \ + \ \gamma^2 \ln\left(\frac{1}{\gamma}\right) \right)$$

This completes the proof of (23), of Lemma 5.1, and of the first claim of the theorem.

▮

5.2 Proof of the second claim of the theorem

First, note that if f is balanced, that is $\mathbb{E}\, f = \frac{1}{2}$, then so is $1-f$, and the second claim of the theorem follows immediately from the first claim.

If $\mathbb{E}\, f \neq \frac{1}{2}$, some additional work is required. We only sketch the argument below, since it is very similar to the proof of the first claim.

Applying Theorem 1.7 to the function $1-f$ gives (cf. (18))

$$Ent\Big(T_\epsilon(1-f)\Big) = Ent\Big(T_{\epsilon_1}(1-h)\Big) \le \lambda_1 \cdot \mathop{\mathbb{E}}_{T, 1\in T} \Big(Ent\Big((1-h) \mid T\Big) - Ent\Big((1-h) \mid \{1\}\Big)\Big) +$$

$$(1-\lambda_1) \cdot \mathop{\mathbb{E}}_{T, 1\notin T} \; Ent\Big((1-h) \mid T\Big) \;\; + \;\; \mathbb{E}(1-h) \cdot \phi\Bigg(Ent\left(\frac{1-h}{\mathbb{E}(1-h)} \;\Big|\; \{1\}\right),\; \epsilon_1\Bigg) \qquad (26)$$

As in the proof of the first claim, we upperbound each of the three summands on the RHS of (26) separately.

Repeating the argument, with the necessary (minor) differences, leads to the same first two bounds:

- $$\mathop{\mathbb{E}}_{T, 1\in T} \Big(Ent\Big((1-h) \mid T\Big) - Ent\Big((1-h) \mid \{1\}\Big)\Big) \;\; \le \;\; O\left(\lambda_1 \cdot \gamma \; + \; \gamma^2 \ln\left(\frac{1}{\gamma}\right)\right)$$
- $$\mathop{\mathbb{E}}_{T, 1\notin T} \; Ent\Big((1-h) \mid T\Big) \;\; \le \;\; O\left(\lambda_1^2 \cdot \gamma \; + \; \lambda_1 \cdot \gamma^2 \ln\left(\frac{1}{\gamma}\right)\right)$$

Indeed, this should not be surprising since, roughly speaking, these two bounds for h are obtained by analysing the behavior of (the squares of) its non-trivial Fourier coefficients, and this is the same for h and for $1-h$.

As to the third summand, we will follow the argument in the proof of Lemma 5.3.

Let $g = \mathbb{E}\left(\frac{1-f}{\mathbb{E}(1-f)} \;\Big|\; \{1\}\right)$. This is a function on a 2-point space $\{0,1\}$, with $g(0) = \rho$ and $g(1) = 2-\rho$, where $\rho = 1 + \frac{(1-\alpha)(1-2\beta)}{1+2\beta}$.

Note that $\rho \le 2 - c\cdot\gamma$, for some absolute constant $c > 0$. Hence, proceeding as in the proof of Lemma 5.3, gives

$$\phi\Bigg(Ent\left(\frac{1-h}{\mathbb{E}(1-h)} \;\Big|\; \{1\}\right),\; \epsilon_1\Bigg) \;\; = \;\; Ent\Big(T_\epsilon g\Big) \;\; \le \;\; \Big(1 - H_2(\epsilon)\Big) - c' \cdot \lambda \cdot \gamma,$$

for an absolute constant $c' > 0$.

Next, recall that in the proof of Lemma 5.3 we show $\phi\left(Ent\left(\frac{h}{\mathbb{E}\, h} \;\Big|\; \{1\}\right),\; \epsilon_1\right) \;\le\; \Big(1 - H_2(\epsilon)\Big) - c_1 \cdot \lambda \cdot \alpha$ for an absolute constant $c_1 > 0$.

Hence

$$\mathbb{E}\, h \cdot \phi\Bigg(Ent\left(\frac{h}{\mathbb{E}\, h}\;\Big|\;\{1\}\right),\ \epsilon_1\Bigg) \ + \ \mathbb{E}(1-h) \cdot \phi\Bigg(Ent\left(\frac{1-h}{\mathbb{E}(1-h)}\;\Big|\;\{1\}\right),\ \epsilon_1\Bigg) \quad \le$$

$$\Big(1 - H_2(\epsilon)\Big) - c_2 \cdot \lambda \cdot \gamma,$$

for an absolute constant $c_2 > 0$.

We can now complete the proof of the second claim of the theorem.

Combining all bounds on the right hand sides of (18) and of (26) above gives

$$Ent\Big(T_\epsilon f\Big) \ + \ Ent\Big(T_\epsilon(1-f)\Big) \quad \le \quad \Big(1 - H_2(\epsilon)\Big) \ - \ \Omega\Big(\lambda \cdot \gamma\Big) \ + \ o_{\lambda,\gamma\to 0}\Big(\lambda \cdot \gamma\Big)$$

Since $\lambda, \gamma \ \le \ O(\delta)$, this implies that for a sufficiently small $\delta > 0$ holds

$$Ent\Big(T_\epsilon f\Big) \ + \ Ent\Big(T_\epsilon(1-f)\Big) \quad \le \quad 1 - H_2(\epsilon)$$

▮

6 Remaining proofs

6.1 Proof of Lemma 1.2

We have, for a boolean function f:

$$I\Big(f(X);Y\Big) \quad = \quad H\Big(f(X)\Big) \ - \ H\Big(f(X)|Y\Big) \quad = \quad H\Big(f(X)\Big) \ - \ \mathop{\mathbb{E}}_{y} H\Big(f(X)|Y=y\Big) \quad =$$

$$H_2\Big(\,\mathbb{E}\, f\Big) \ - \ \mathop{\mathbb{E}}_{y} H_2\Big(\,(T_\epsilon f)\,(y)\Big)$$

We have $H_2\left(\mathbb{E}\, f\right) = \mathbb{E}\, f \log \frac{1}{\mathbb{E}\, f} + (1 - \mathbb{E}\, f) \log \frac{1}{1-\mathbb{E}\, f}$.

We also have (all the logarithms are binary)

$$\mathop{\mathbb{E}}_{y} H_2\Big(\,(T_\epsilon f)\,(y)\Big) \quad = \quad \mathop{\mathbb{E}}_{y}\left((T_\epsilon f)\,(y) \log \frac{1}{(T_\epsilon f)\,(y)} \ + \ \Big(1 - (T_\epsilon f)\,(y)\Big) \log \frac{1}{1 - (T_\epsilon f)\,(y)}\right) \quad =$$

$$-\Bigg(Ent\Big(T_\epsilon f\Big) \ + \ \mathbb{E}\, T_\epsilon f \log \mathbb{E}\, T_\epsilon f\Bigg) \ - \ \Bigg(Ent\Big(T_\epsilon(1-f)\Big) \ + \ \mathbb{E}\, T_\epsilon(1-f) \log \mathbb{E}\, T_\epsilon(1-f)\Bigg) \quad =$$

$$-\Bigg(Ent\Big(T_\epsilon f\Big) + Ent\Big(T_\epsilon(1-f)\Big)\Bigg) \ + \ \mathbb{E}\, f \log \frac{1}{\mathbb{E}\, f} \ + \ (1 - \mathbb{E}\, f) \log \frac{1}{1 - \mathbb{E}\, f}$$

In the last step we have used the fact $\mathbb{E}\, T_\epsilon g \ = \ \mathbb{E}\, g$ for any function g. The claim of the lemma follows. ▮

6.2 Proof of Corollary 1.10

Applying Corollary 1.9 to the functions f and $1-f$, we obtain, by Lemma 1.2:

$$I\Big(f(X);Y\Big) \quad = \quad Ent\Big(T_\epsilon f\Big) \ + \ Ent\Big(T_\epsilon(1-f)\Big) \ \le \ \mathop{\mathbb{E}}_{T}\Big(Ent\Big(f \mid T\Big) \ + \ Ent\Big((1-f) \mid T\Big)\Big)$$

To conclude the proof of the corollary, it suffices to show that for any $T \subseteq [n]$ holds

$$Ent\Big(f \mid T\Big) \ + \ Ent\Big((1-f) \mid T\Big) \quad = \quad I\Big(f(X);\ \{X_i\}_{i\in T}\Big)$$

To see this, we proceed exactly as in the proof of Lemma 1.2, observing that, by definition,

$$Pr\Big\{f(X)=1 \ \Big| \ \{X_i\}_{i\in T}\Big\} \quad = \quad \mathbb{E}\Big(f \mid T\Big)$$

Here we interpret both sides as functions of $\{x_i\}$, $i \in T$.

▮

6.3 Proof of Theorem 1.12

The proof of this theorem is very similar to that of Theorem 1.7 and uses the notation and some of the results from that proof.

As in the proof of Lemma 4.3, our starting point is the chain rule for noisy entropy (7), which states that for any permutation $\sigma \in S_n$ the noisy entropy $Ent\Big(T_\epsilon f\Big)$ is bounded from above by

$$\sum_{i=1}^{n} \phi\Big(Ent\Big(f \mid \{\sigma(i)\}\Big) \ + \ I_{T_{\epsilon_{\{\sigma(1),\ldots,\sigma(i-1)\}}} f}\ \Big(\Big\{\sigma(1),\ldots,\sigma(i-1)\Big\},\ \sigma(i)\Big)\Big)$$

Averaging over $\sigma \in S_n$ and using transitivity of action of the symmetric group and concavity of ϕ, this is at most

$$\sum_{k=0}^{n-1} \phi\left(\mathop{\mathbb{E}}_{i\in[n]} Ent\Big(f \mid i\Big) \ + \ b_k\right) \quad \text{where} \quad b_k \quad = \quad \mathop{\mathbb{E}}_{A,m} T_{\epsilon_A f}\ \Big(A, m\Big)$$

and the expectation is over all $A \subseteq [n]$ of cardinality k and $m \notin A$. (In particular, we set $b_0 = 0$). Using the concavity of ϕ again, this is at most

$$n \cdot \phi\left(\mathop{\mathbb{E}}_{i\in[n]} Ent\Big(f \mid i\Big) \ + \ \frac{1}{n} \cdot \sum_{k=0}^{n-1} b_k\right)$$

The analysis in the proof of Theorem 1.7 shows that if T is a random subset of $[n]$ generated by sampling each element $i \in [n]$ independently with probability λ then

$$\sum_{k=0}^{n-1} b_k \quad = \quad \frac{1}{\lambda} \cdot \sum_{s=1}^{n-1} w_s \cdot t_s \quad = \quad \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{T} \Big(Ent\Big(f \mid T\Big) \;-\; \sum_{i \in T} Ent\Big(f \mid i\Big)\Big) \quad =$$

$$\frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{T} Ent\Big(f \mid T\Big) \;-\; n \cdot \mathop{\mathbb{E}}_{i \in [n]} \; Ent\Big(f \mid i\Big)$$

Substituting and simplifying, we get, setting $t = \lambda n$,

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad n \cdot \phi \left(\frac{\mathbb{E}_T \, Ent\Big(f \mid T\Big)}{t} \right)$$

which is the claim of the theorem. ▮

6.3.1 Proof of (6)

Let f be the distribution of X multiplied by 2^n. Then $\mathbb{E} f = 1$, and Theorem 1.12 can be applied.

By Section 1.1.1 and (5), we have $Ent\Big(T_\epsilon f\Big) \;=\; n - H\Big(X \oplus Z\Big)$ and $Ent\Big(f \mid T\Big) \;=\; |T| - H\Big(\{X_i\}_{i \in T}\Big)$.

We also recall $\phi\Big(x, \epsilon\Big) \;=\; 1 - H_2\Big(\epsilon + (1-2\epsilon) \cdot H_2^{-1}(1-x)\Big)$.

Substituting in the claim of the theorem, and simplifying, gives

$$H\Big(X \oplus Z\Big) \quad \ge \quad n \cdot H_2 \left(\epsilon \;+\; (1-2\epsilon) \cdot H_2^{-1} \left(\frac{\mathbb{E}_T \, H\Big(\{X_i\}_{i \in T}\Big)}{t} \right)\right)$$

which is the claim of (6).

▮

6.4 Proof of Theorem 1.15

Let δ be the constant in the theorem. We will assume that δ is sufficiently small.

Let ϵ be a noise parameter, such that $(1-2\epsilon)^2 \le \delta$. Denote $\lambda = (1-2\epsilon)^2$.

It is known (see [7]) that for any boolean function f holds $I\left(f(X); Y\right) \;\le\; \lambda \cdot H_2\left(\mathbb{E} f\right)$. This immediately implies the validity of Conjecture 1.1 for boolean functions with expectation lying in $[0, c] \;\cup\; [1-c, 1]$, for some absolute constant $0 < c < 1/2$.

In addition, we may assume, by symmetry, that $\mathbb{E}\, f \le 1/2$. Combining these two observations, it remains to consider the case

$$c \quad \le \quad \mathbb{E}\, f \quad \le \quad 1/2 \tag{27}$$

Let f be a boolean function satisfying (27) with $I\left(f(X); Y\right) \ge 1 - H_2(\epsilon)$. This is the same as $Ent\left(T_\epsilon f\right) + Ent\left(T_\epsilon (1-f)\right) \ge 1 - H_2(\epsilon)$.

At this point, we need a technical lemma.

Lemma 6.1: *For any nonnegative non-zero function f holds*[5]

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad \frac{1}{\mathbb{E}\, f} \cdot \left(\frac{1}{2\ln 2} \cdot \sum_{k=1}^n \widehat{f}^2(\{k\})\right) \cdot \lambda \; + \; O_{\lambda \to 0}\left(\frac{\mathbb{E}\, f^2}{\mathbb{E}\, f} \cdot \lambda^{4/3}\right) \; + \; O_{\lambda \to 0}\left(\frac{\mathbb{E}^2\, f^2}{\mathbb{E}^3\, f} \cdot \lambda^2\right)$$

We will now proceed with the proof of the theorem, and prove the lemma below.

Applying Lemma 6.1 to functions f and $1-f$ and taking into account (27) gives

$$Ent\Big(T_\epsilon f\Big) \; + \; Ent\Big(T_\epsilon (1-f)\Big) \quad \le \quad \frac{1}{\mathbb{E}\, f\Big(1 - \mathbb{E}\, f\Big)} \cdot \left(\frac{1}{2\ln 2} \cdot \sum_{k=1}^n \widehat{f}^2(\{k\})\right) \cdot \lambda \; + \; O_{\lambda \to 0}\Big(\lambda^{4/3}\Big)$$

Combining the two inequalities for $Ent\left(T_\epsilon f\right) + Ent\left(T_\epsilon (1-f)\right)$, recalling $1 - H_2(\epsilon) \;\ge\; \frac{\lambda}{2\ln 2}$, and using (27), we get

$$\sum_{k=1}^n \widehat{f}^2(\{k\}) \quad \ge \quad \mathbb{E}\, f \cdot \Big(1 - \mathbb{E}\, f\Big) \; - \; O\left(\lambda^{1/3}\right)$$

For a boolean function $f : \{0,1\}^n \to \{0,1\}$ holds $\mathbb{E}\, f^2 = \mathbb{E}\, f$, and consequently $\sum_{S \ne \emptyset} \widehat{f}^2(S) \;=\; \mathbb{E}\, f(1 - \mathbb{E}\, f)$. Hence we have

$$\sum_{|S| \ge 2} \widehat{f}^2(S) \le O\left(\lambda^{1/3}\right)$$

Let $g = 2f - 1$. Then $g :\; \{0,1\}^n \to \{-1,1\}$, and $\sum_{|S|\ge 2} \widehat{g}^2(S) = 4 \cdot \sum_{|S| \ge 2} \widehat{f}^2(S) \le O\left(\lambda^{1/3}\right)$. Note also that (27) implies $2c - 1 \le \widehat{g}(0) = 2\,\mathbb{E}\, g - 1 \le 0$.

Hence, assuming λ is sufficiently small, Theorem 5.5 implies that $|\,\mathbb{E}\, g| \le O\left(\lambda^{1/3} \cdot \sqrt{\ln \frac{1}{\lambda}}\right)$, and there exists an index $1 \le k \le n$ such that $\widehat{g}^2(\{k\}) \ge 1 - O\left(\lambda^{1/3}\right)$.

[5] Asymptotic notation hides absolute constants independent of the remaining parameters.

This means that

$$\frac{1}{2} \; - \; O\left(\lambda^{1/3} \cdot \sqrt{\ln \frac{1}{\lambda}}\right) \quad \le \quad \mathbb{E} f \quad \le \quad \frac{1}{2} \quad \text{and} \quad |\widehat{f}(\{k\})| \quad \ge \quad \frac{1}{2} \; - \; O\left(\lambda^{1/3}\right)$$

If λ is sufficiently small, f satisfies the conditions of Theorem 1.14. By the second claim of this theorem,

$$Ent\,(T_\epsilon f) \;\; + \;\; Ent\,(T_\epsilon(1-f)) \quad \le \quad 1 \; - \; H_2(\epsilon),$$

completing the proof of Theorem 1.15.

▮

6.4.1 Proof of Lemma 6.1

The argument below is a slight extension of an argument in [14].

In the following, we may and will assume, by homogeneity, that $\mathbb{E} f = 1$.

Let us introduce some notation. For $x \in \{0,1\}^n$, let x^c be the complement of x, that is the element of $\{0,1\}^n$ with $x^c_i = 1 - x_i$ for all $1 \le i \le n$.

For a nonnegative function g on $\{0,1\}^n$, let g_0 be the 'even' part of g defined by $g_0(x) \;=\; (g(x) + g\,(x^c))\,/2$, and let $g_1 \;=\; g - g_0$ be the 'odd' part of g. By definition, $g_0(x) \;=\; g_0\,(x^c)$ and $g_1(x) \;=\; -g_0\,(x^c)$.
Note also that $|g_1| \le g_0$.

We will need the following well-known (and easy to verify) fact:

$$g_0 \quad = \sum_{|S|\ even} \widehat{g}(S) \cdot W_S \quad \text{and} \quad g_1 \quad = \sum_{|S|\ odd} \widehat{g}(S) \cdot W_S$$

We start with an auxiliary claim.

Lemma 6.2: *For a function g with $\mathbb{E}\, g = 1$ holds*

$$Ent\,(g) \quad = \quad Ent\Big(g_0\Big) \;\; + \;\; \mathop{\mathbb{E}}_{x} \; g_0(x) \cdot \left(1 - H_2\left(\frac{1 - |g_1(x)|/g_0(x)}{2}\right)\right)$$

Here for x such that $g_0(x) \;=\; g_1(x) \;=\; 0$, the expression $g_0(x) \cdot \left(1 - H_2\left(\frac{1-|g_1(x)|/g_0(x)}{2}\right)\right)$ is interpreted as 0.

Proof: We have

$$Ent\Big(g\Big) \quad = \quad \mathop{\mathbb{E}}_{x} g(x) \log g(x) \quad = \frac{1}{2} \cdot \mathop{\mathbb{E}}_{x} \Big(g(x) \log g(x) \;\; + \;\; g\,(x^c) \log\,(x^c)\Big) \quad =$$

$$\frac{1}{2} \cdot \mathop{\mathbb{E}}_{x} \left(\Big(g_0(x) + g_1(x)\Big) \cdot \log \Big(g_0(x) + g_1(x)\Big) \quad + \quad \Big(g_0(x) - g_1(x)\Big) \cdot \log \Big(g_0(x) - g_1(x)\Big) \right)$$

It is easy to verify that for any $0 \le b \le a$ holds $1/2 \cdot \Big((a+b)\log(a+b) + (a-b)\log(a-b)\Big) \;=\; a \log a + a \cdot \left(1 - H_2\left(\frac{1-b/a}{2}\right)\right)$, where the last expression should be interpreted as 0 for $a = b = 0$.

Using this identity with $a = g_0(x)$ and $b = g_1(x)$ gives the claim of the lemma. ∎

Next, as in [14], we upper bound $1 - H_2\left(\frac{1-x}{2}\right)$ by $\frac{1}{2\ln 2} \cdot x^2 + \left(1 - \frac{1}{2\ln 2}\right) \cdot x^4$.

Substituting this bound in the claim of Lemma 6.2 gives

$$Ent\,(g) \quad \le \quad Ent\Big(g_0\Big) \quad + \quad \frac{1}{2\ln 2} \cdot \mathop{\mathbb{E}}_{x} \frac{g_1^2(x)}{g_0(x)} \quad + \quad \left(1 - \frac{1}{2\ln 2}\right) \cdot \mathop{\mathbb{E}}_{x} \frac{g_1^4(x)}{g_0^3(x)}$$

Let $g = T_\epsilon f$. It is easy to verify $\Big(T_\epsilon f\Big)_i = T_\epsilon \Big(f_i\Big)$ for any function f and $i = 0, 1$. Consequently, we get the bound

$$Ent\Big(T_\epsilon f\Big) \quad \le \quad Ent\Big(T_\epsilon f_0\Big) \;+\; \frac{1}{2\ln 2} \cdot \mathop{\mathbb{E}}_{x} \frac{\Big(T_\epsilon f_1(x)\Big)^2}{T_\epsilon f_0(x)} \;+\; \left(1 - \frac{1}{2\ln 2}\right) \cdot \mathop{\mathbb{E}}_{x} \frac{\Big(T_\epsilon f_1(x)\Big)^4}{\Big(T_\epsilon f_0(x)\Big)^3}$$

We upperbound each of the summands on the RHS separately.

1. The first summand. Note that $\mathbb{E}\, T_\epsilon f_0 = \mathbb{E}\, f_0 = \mathbb{E}\, f = 1$. Recall also (see e.g., [14]) that for any function g on $\{0,1\}^n$ holds $T_\epsilon g \;=\; \sum_S \lambda^{|S|/2} \widehat{g}(S) \cdot W_S$.

   Hence, by Lemma 5.4,

$$Ent\Big(T_\epsilon f_0\Big) \quad \le \quad O\Big(\sum_S |S| \cdot \widehat{T_\epsilon f_0}^2(S)\Big) \quad = \quad O\Big(\sum_S |S| \lambda^{|S|} \widehat{f_0}^2(S)\Big) \quad =$$

$$O\Big(\sum_{|S|\ even} |S| \lambda^{|S|} \widehat{f}^2(S)\Big) \quad = \quad O\Big(\mathbb{E}\, f^2 \cdot \lambda^2\Big)$$

2. The second summand.

   First, we argue that $T_\epsilon f_0$ is bounded away from 0 with high probability. Recall that $\mathbb{E}\, T_\epsilon f_0 \;=\; 1$, and note that $Var\Big(T_\epsilon f_0\Big) \;=\; \sum_{S \ne 0} \widehat{T_\epsilon f_0}^2(S) \;=\; O\Big(\lambda^2 \cdot \mathbb{E}\, f^2\Big)$. Hence, by Chebyshev's inequality, for any $0 \le \alpha < 1$ holds

$$Pr\Big\{T_\epsilon f_0 \le \alpha\Big\} \quad \le \quad O\left(\frac{\lambda^2 \cdot \mathbb{E}\, f^2}{(1-\alpha)^2}\right) \tag{28}$$

   Second, recall that for any x holds $|T_\epsilon f_1(x)| \le T_\epsilon f_0(x)$, and hence $\frac{(T_\epsilon f_1(x))^2}{T_\epsilon f_0(x)} \;\le\; T_\epsilon f_0(x)$.

Therefore, taking $\alpha = 1 - \lambda^{1/3}$ in (28), we have $\mathbb{E}_x \frac{(T_\epsilon f_1(x))^2}{T_\epsilon f_0(x)}$ bounded from above by

$$Pr\Big\{f_0 \ \le \ 1 - \lambda^{1/3}\Big\} \cdot \alpha \ + \ \Big(1 + O\left(\lambda^{1/3}\right)\Big) \cdot \mathop{\mathbb{E}}_x \Big(T_\epsilon f_1(x)\Big)^2$$

Recalling that $T_\epsilon f_1 = \sum_{|S| \ odd} \lambda^{|S|/2} \widehat{f}(S) \cdot W_S$, this is at most

$$O\Big(\mathbb{E} f^2 \cdot \lambda^{4/3}\Big) \ + \ \Big(1 + O\left(\lambda^{1/3}\right)\Big) \cdot \left(\left(\sum_{k=1}^n \widehat{f}^2(\{k\})\right) \cdot \lambda \ + \ O\Big(\mathbb{E} f^2 \cdot \lambda^3\Big)\right) \ =$$

$$\left(\sum_{k=1}^n \widehat{f}^2(\{k\})\right) \cdot \lambda \ + \ O\Big(\mathbb{E} f^2 \cdot \lambda^{4/3}\Big)$$

3. The third summand. Note that $\mathbb{E} f_1 = 0$. Hence, as in Lemma 1 in [14] (where the requirement on f to be boolean does not seem to be necessary) we have, for a sufficiently small λ, that

$$\mathop{\mathbb{E}}_x \Big(T_\epsilon f_1(x)\Big)^4 \ \le \ O\left(\Big(\mathop{\mathbb{E}}_x f_1^2(x)\Big)^2 \cdot \lambda^2\right) \ = \ O\left(\mathbb{E}^2 f^2 \cdot \lambda^2\right)$$

We can now upperbound the third summand using the Chebyshev inequality, as above. Taking $\alpha = 1/2$ in (28) upperbounds $\mathbb{E}_x \frac{(T_\epsilon f_1(x))^4}{(T_\epsilon f_0(x))^3}$ by

$$O\left(\mathbb{E} f^2 \cdot \lambda^2\right) \cdot \alpha \ + \ O\left(\mathbb{E}^2 f^2 \cdot \lambda^2\right) \ = \ O\left(\mathbb{E} f^2 \cdot \lambda^2\right) \ + \ O\left(\mathbb{E}^2 f^2 \cdot \lambda^2\right)$$

Combining these estimates leads to the claim of Lemma 6.1. ▮

▮

Acknowledgments

We are grateful to Yuval Kochman, Or Ordentlich, and Yury Polyanskiy for many very helpful conversations and valuable remarks. We also thank Venkat Chandar for valuable remarks.

References

[1] N. Alon and J. Spencer, **The Probabilistic Method**, 3rd ed. Hoboken, NJ: Wiley, 2008.

[2] R. Ahlswede and P. Gacs, *Spreading of sets in product spaces and hypercontraction of the Markov operator*, Ann. Probab., vol. 4, no. 6, pp. 925-939, 1976.

[3] V. Chandar, personal communication, 2014.

[4] E.Friedgut, G. Kalai, and A. Naor, *Boolean functions whose Fourier transform is concentrated on the first two levels*, Advances in Applied Mathematics, vol. 29, 3, pp. 427-437, 2002.

[5] T. S. Han, *Nonnegative entropy measures of multivariate symmetric correlations*, Inform. Contr. 36, pp. 133-156, 1978.

[6] J. Jendrej, K. Oleszkiewicz, and J. O. Wojtaszczyk, *On some extensions of the FKN theorem*, Theory of Computing, vol. 11 (18), pp. 445-469, 2015.

[7] T. Courtade and G. Kumar, *Which boolean functions maximize mutual information on noisy inputs?* IEEE Transactions on Information Theory, vol. 60, no. 8, pp. 4515-4525, 2014.

[8] M. Ledoux, **Concentration of measure and logarithmic Sobolev inequaities**, 1997.

[9] E. Mossel and S. Sachdeva, personal communication, 2014.

[10] A. W. Marshall, and I. Olkin, **Inequalities: Theory of Majorization and Its Applications**, Academic, New York, 1979.

[11] R. O'Donnel, **Analysis of Boolean functions**, Cambridge U.P., 2014.

[12] O. Ordentlich, personal communication, 2015.

[13] O. Ordentlich, *On the entropy rate of binary hidden Markov processes*, 2016, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1601.06453.

[14] O. Ordentlich, O. Shayevitz, and O.Weinstein, *An Improved Upper Bound for the Most Informative Boolean Function Conjecture*, 2015, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1505.05794.

[15] Y. Polyanskiy, personal communication, 2015.

[16] Y. Polyanskiy and Y. Wu, *Dissipation of information in channels with input constraints*, 2014, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1405.3629.

[17] Y. Polyanskiy and Y. Wu, *Strong data-processing inequalities for channels and Bayesian networks*, 2016, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1508.06025.

[18] S. Sachdeva, A. Samorodnitsky, and I. Shahaf, *On conjectures of Kumar and Courtade*, manuscript, 2014.

[19] A. Samorodnitsky, mansucript, 2015.

[20] M. Talagrand, *How much are increasing sets positively correlated?*, Combinatorica 16.2, 243-258, 1996.

[21] A. D. Wyner and J. Ziv, *A theorem on the entropy of certain binary sequences and applications: Part I*, IEEE Trans. Inform. Theory, vol. 19, no. 6, pp. 769-772, 1973.