ON THE BIT EXTRACTION PROBLEM

Joel Friedman

CS-TR-357-91

December 1991

On the Bit Extraction Problem

Joel Friedman*
Department of Computer Science
Princeton University
Princeton, NJ 08544

December 4, 1991

Abstract

Consider a coloring of the n-dimensional Boolean cube with $c = 2^s$ colors in such a way that every k-dimensional subcube is equicolored, i.e. each color occurs the same number of times. We show that for such a coloring we necessarily have $(k-1)/n \geq \theta_c = (c/2-1)/(c-1)$. This resolves the "bit extraction" or "t-resilient functions" problem in many cases, such as $c-1|n$, proving that XOR type colorings are optimal. We also study the problem of finding almost equicolored colorings when $(k-1)/n < \theta_c$, and of classifying all optimal colorings.

1 Introduction

The bit extraction, aka t-resilient functions (see [CFG+85]), aka privacy amplification problem (see [BBR88],[Bra89]) is the following. The vertices of the Boolean cube, $\mathbf{B}^n = \{-1,1\}^n$ are to be colored with $c = 2^s$ colors such that every *k-dimensional subcube* is *equicolored.* Given n and s, what is the smallest value of k for which this possible? Here, by a k-dimensional subcube we mean a subset of $\mathbf{B}^n$ determined by fixing the values of some $n-k$ coordinates on $\mathbf{B}^n$; we denote the set of all such subcubes by $\mathcal{H}_k$. By equicolored we mean that every color appears the same number of times in the subcube, i.e. $2^k/c$ times.

This question comes up in various contexts, such as in maintaining privacy (see [BBR88]) and in deriving unbiased random bits from a source of bits with some unknown subset of the bits being biased (see [CFG+85]). We refer to such a coloring as a $(c; n, k)$-coloring and write $\kappa(c, n)$ for the smallest value of k achievable for a given n and c. In this paper we are primarily interested in viewing c as small or fixed, studying

*The author also wishes to acknowledge the National Science Foundation for supporting this research in part under a PYI grant, CCR–8858788.

κ as a function of n. We will study this problem and the problem of constructing colorings which are approximately equicolorable.

The optimal known colorings are constructed as follows. If we identify the colors with $\mathbf{B}^s$, then the coloring determines s Boolean functions $f_1, \ldots, f_s$, $f_i: \mathbf{B}^n \to \{0,1\}$; conversely, any such s-tuple of Boolean functions determines a coloring. By an *XOR coloring* we mean one where each f_i is the XOR (exclusive-or, i.e. multiplication in $\{-1,1\}$) of some subset of the coordinates, $\{x_1, \ldots, x_n\}$, of $\mathbf{B}^n$. It is easy to determine, to within $O(c)$, how small a k can be achieved via such colorings. For example, for n divisible by $c-1$, it is easy to see that we can achieve $k = n\theta_c + 1$ and no smaller value, where $\theta_c = (c/2-1)/(c-1)$. It follows that for fixed $c = 2^s$ and any $\epsilon > 0$ one has $\kappa(c,n)/n \leq \theta_c + \epsilon$ for n sufficiently large. There are no known values of c, n for which one can provably do better than XOR colorings.

It has been conjectured that XOR colorings are optimal for any $n, c = 2^s$. This was proven for all n with $s = 2$ in [CFG+85] (and is obvious for $s = 1$). Here we prove the following:

Theorem 1.1 *For any $(c; n, k)$ coloring we have $(k-1)/n \geq \theta_c$.*

Corollary 1.2 *For $c-1|n$, $n > 0$, XOR colorings are optimal. For any n, c, XOR colorings are within $(c-2)/2$ of optimal.*

It is a non-trivial problem to determine exactly what value of k one can achieve by XOR colorings when $c-1$ does not divide n, but one can do so in certain cases to obtain other bounds as a corollary. For example, one also gets:

Proposition 1.3 *For $n \geq 2$ and $\equiv -2, -1, 0, 1, c/2-1, c/2, c/2+1 \pmod{c-1}$, XOR colorings can achieve $1 + \lceil n\theta_c \rceil$ and hence are optimal. For $n > 2$ and $\equiv 2 \pmod{c-1}$ with $c \geq 8$, optimal XOR colorings achieve $2 + \lceil n\theta_c \rceil$, and hence yield k's at worst within 1 of optimal. For $c = 8$ the former and latter congruences hold according to whether $n \not\equiv 2 \pmod 7$ or not. For any n, c, XOR colorings are within $c/4$ of optimal.*

In particular, $n \equiv 2 \pmod 7$ is the simplest case in which we don't know if XOR colorings are optimal.

We are also interested in two related problems:

Problem 1.4 *For $k < \kappa(c,n)$, how close (in various metrics involving $\mathcal{H}_k$) to equicolorable can be achieved by a c-coloring of $\mathbf{B}^n$?*

Problem 1.5 *Classify all optimal colorings or s-tuples of functions, at least for $c-1|n$. Are there optimal colorings not obtainable as XOR type colorings?*

The author knows of no optimal colorings other than XOR colorings.

In this paper we study problem 1.4 for the metric $L^2(\mathcal{H}_k)$ (defined in section 3). We give a lower bound on the distance to "equicolor" one can achieve for $k < \kappa(c,n)$. One remarkable fact is:

Theorem 1.6 *For $c-1|n$, the optimal XOR coloring is closest to equicolorable in $L^2(\mathcal{H}_k)$ for any k.*

In general, XOR colorings are not closest to equicolor for many important metrics. For example, consider the case $n = c = 4, k = 2$. Fix an optimal XOR coloring $\gamma: \mathbf{B}^3 \to \mathbf{B}^2$, and let $\delta: \mathbf{B}^4 \to \mathbf{B}^2$ be the coloring defined on $\mathbf{B}^4$ regarded as $\mathbf{B}^3 \times \mathbf{B}$ via $\delta(x, 1) = \gamma(x)$, and $\delta(x, -1) = \sigma(\gamma(x))$ for any permutation, σ, of $\mathbf{B}^2$ of order 4. Every $H \in \mathcal{H}_2$ contains at least three colors for δ, while for any XOR coloring there are H's which contain only two colors. In particular, δ is closer to equicolor than any XOR coloring in the $\mathrm{RP}(\mathcal{H}_k)$ metric defined in section 3. In general, for k one less than what can be achieved by the best XOR coloring for fixed c, n, the best XOR colorings are equicolored on almost all H's, but on the other H's they avoid half the colors! It would seem that one could do much better by spreading out badness of these H's to a much larger subset of $\mathcal{H}_k$. It would be interesting to know (or have a reasonable conjecture) about the closest to equicolor colorings for various sup-type norms, such as $\mathrm{RP}(\mathcal{H}_k)$ or $L^\infty(\mathcal{H}_k)$.

In the proof that XOR colorings are optimal for $c = 4$ of [CFG+85], one produces from an optimal coloring f_1, f_2 two subsets of variables X_1, X_2 whose XOR's yield an optimal coloring (any X_1, X_2 with the Fourier coefficient of $f_1, f_2, f_1 + f_2$ non-vanishing at, respectively, $X_1, X_2, X_1 + X_2$ will do). This proof does not directly generalize because of the possibility of cancellation of Fourier coefficients in computing a convolution, and theorem 1.1 somehow precludes very bad cancellation. It would be nice to find a generalization of the method in [CFG+85], understanding precisely how much cancellation can occur in such convolutions. In our paper there is no explicit reference to Fourier coefficients of the f_i's. They occur only implicitly, in that the eigenspace of the adjacency matrix of $\mathbf{B}^n$ corresponding to the eigenvalue $n - 2r$ is precisely the ($\mathbf{R}$-linear) span of the collection of all XOR's of r variables.

In section 2 we give a short proof of theorem 1.1 and prove some facts about the optimal XOR colorings, proving proposition 1.3. In section 3 study the approximate equicoloring problem for L^2. To do so we study, for a subset $C \subset \mathbf{B}^n$, the statistics p_i defined to be the probability that a random vertex at distance i from a random vertex of C is again a vertex of C. We refer to this as the *profile* of C. The profile has many intriguing properties, and we comment on some of them in section 4. It seems important to understand the profile better, for example, in order to study approximately equicolored colorings in different norms such as the sup-type norms— while the method of section 3 does not directly generalize to the sup-type or to other L^p norms, there is a natural generalization of the profile which determines distance to equicolor in L^p for p an even integer; understanding the profile and "higher-order" profiles might shed light on the L^p problem for larger p's (and maybe, thus, $p = \infty$). In section 5 we comment more on problem 1.5, giving a simple geometric characterization of all XOR colorings.

The author wishes to thank Kai Li, Bernard Chazelle, and Avi Wigderson for useful comments and discussions.

2 The Bit Extraction Problem

We begin by proving theorem 1.1. We do this via a somewhat stronger statement. We say that a subset $C \subset \mathbf{B}^n$ is $1/c$ dense in $\mathcal{H}_k$ if

$$\frac{|C \cap H|}{|H|} = \frac{1}{c} \qquad \forall H \in \mathcal{H}_k.$$

Theorem 2.1 *If there exists a $1/c$ dense in $\mathcal{H}_k$ subset of $\mathbf{B}^n$, then $(k-1)/n \geq \theta_c$.*

Theorem 1.1 follows by taking C to be the set of vertices of any fixed color of a $(c; n, k)$ coloring.

Proof Consider the adjacency matrix, A of the Boolean cube, and χ_C, the characteristic function of C in $\mathbf{B}^n$. Clearly $(A\chi_C, \chi_C) \geq 0$. On the other hand, the eigenvalues of A are $n - 2r$ with $r = 0, 1, \ldots n$, and the corresponding eigenspaces, E_r, are just the spans of all XOR's of r variables. If v_r is the projection of χ_C onto E_r, setting $\mu_r = |v_r|^2/|\chi_c|^2$, we have

$$\sum_{r=0}^{n} \mu_r = \frac{\sum |v_r|^2}{|\chi_C|^2} = 1, \qquad (A\chi_C, \chi_C) = \sum_{r=0}^{n} (n-2r)|v_r|^2 = |\chi_C|^2 \sum_{r=0}^{n} (n-2r)\mu_r.$$

But E_0 corresponds to the trivial eigenvector, $(1, \ldots, 1)$, and so $\mu_0 = |C|/n = 1/c$; also, the assumption that C is $1/c$ dense in $\mathcal{H}_k$ means that $v_1 = \cdots = v_{n-k} = 0$ and hence $\mu_1 = \cdots = \mu_{n-k} = 0$. Hence

$$0 \leq n\mu_0 + \sum_{r=n-k+1}^{n} (n-2r)\mu_r \leq n\mu_0 + (2k-2-n) \sum_{r=n-k+1}^{n} \mu_r = n\frac{1}{c} + (2k-2-n)\frac{c-1}{c},$$

and so

$$2(k-1)\frac{c-1}{c} \geq n\frac{c-2}{c},$$

which is the desired result.

□

We will now work out the optimal XOR colorings for some special cases to deduce corollary 1.2 and proposition 1.3. Let $\kappa_{\mathrm{XOR}}(c, n)$ denote the optimal k achievable by XOR colorings. We will have occasion to use the easy:

Proposition 2.2 *If there exists an $(c; n, k)$-coloring, then for any integer $r > 0$ there exist $(c; n + r, k + r)$ and $(c; nr, (k-1)r + 1)$ colorings, and for any positive integer $r \leq k$ there exists a $(c; n - r, k)$ coloring.*

Proof If $\gamma: \mathbf{B}^n \to \mathbf{B}^s$ is such a coloring, then any projection $\pi: \mathbf{B}^{n+r} \to \mathbf{B}^n$ yields an $(c; n+r, k+r)$ coloring, $\gamma \circ \pi$. Similarly, $\gamma^r: \mathbf{B}_{nr} \to (\mathbf{B}_k)^r$ defined in the obvious way, followed by $g: (\mathbf{B}_k)^r \to \mathbf{B}_k$ given by bitwise XORing yields an $(c; nr, (k-1)r+1)$ coloring. Similarly, restricting γ to any $n-r$ dimensional subcube of $\mathbf{B}^n$ yields a $(c; n-r, k)$ coloring.

The above proposition implies that $\kappa(c, n+1) - \kappa(c, n)$ is either 0 or 1, for all c, n, and similarly for κ_{XOR} (since the new colorings produced from γ are XOR colorings if γ is).

To analyze the optimal XOR colorings, first recall that in general $f_1, \ldots, f_s$ yield an $\mathcal{H}_k$ equicolored coloring iff all XOR's of a subsets of $\{f_1, \ldots, f_s\}$ yields a function which is half 1, half -1 on every $H \in \mathcal{H}_k$ (see, for example, [CFG+85]; this is just to say that the standard $2^s \times 2^s$ Hadamard matrix is invertible). So for a subset $T \subset S = \{1, \ldots, s\}$, consider the XOR of the f_i with $i \in T$, which we denote f_T. If the f_i are XOR's of a subset the variables $X = \{x_1, \ldots, x_n\}$, then so is each f_T. Furthermore, an XOR of the variables is half 1, half -1 on $\mathcal{H}_k$ iff it is the XOR of at least $n - k + 1$ variables. This reduces the analysis of optimal XOR colorings to a question about the possible Venn diagrams of s subsets, $X_1, \ldots, X_s$, of X (or to a question about error correcting codes, as in [CFG+85]).

Namely, for a non-empty $T \subset S$, consider the size of the corresponding component of the Venn diagram on the X_i's,

$$I_T = |\left(\cap_{i \in T} X_i\right) \cap \left(\cap_{i \notin T} \overline{X}_i\right)|,$$

where $\overline{X}_i$ is the complement of X_i in X. The X_i's correspond to an $\mathcal{H}_k$ equicolored coloring iff for all $U \subset S$,

$$\sum_{|T \cap U| \equiv 1 \pmod 2} I_T \geq n - (k-1). \tag{2.1}$$

Furthermore if there exist non-negative, integral I_T satisfying the above equation with $n \geq k$, then clearly there exists a $(c; n, k)$ XOR coloring.

Summing the above over all U shows that

$$\kappa_{\text{XOR}}(c, n) \geq 1 + n\theta_c, \tag{2.2}$$

and if equality holds then each of the inequalities of equation 2.1 holds with equality; the invertibiliy of the standard Hadamard matrix implies that equality holding in all the above inequalities necessitates $I_T = n/(c-1)$. So for $c-1|n$, any choice of X_i with $I_T = n/(c-1)$ for all T yields an optimal coloring, and any optimal coloring occurs in this way. For $c-1|n$ we use the term *balanced* XOR coloring for any optimal XOR coloring to emphasise the fact that all I_T's are equal. Furthermore we have

Proposition 2.3 *For $n \geq 2$ and $\equiv -2, -1, 0, 1, c/2-1, c/2, c/2+1 \pmod{c-1}$, $\kappa_{\text{XOR}}(c, n) = 1 + \lceil n\theta_c \rceil$, and thus optimal colorings can be achieved by XOR colorings in these cases.*

Proof (In analyzing $n \equiv c/2-1, c-3 \pmod{c-1}$ we are assuming $c \geq 8$ in what follows, so that, e.g., $\kappa_{\mathrm{XOR}}(c, c-3)$ exists, i.e. $\mathbf{B}^{c-3} \geq c$; we needn't worry about these cases for $c = 4, 2$.) If $\kappa_{\mathrm{XOR}}(c,n) = 1 + \lceil n\theta_c \rceil$ for some $n = n_0$, then it holds for all $n = n_0 + r(c-1)$ with r any positive integer (by adding r to each I_T). So it suffices to check the above in the cases $n = c/2-1, c/2, c/2+1, c-3, c-2, c-1, c$, and to note that $1 + \lceil n\theta_c \rceil \leq n$ for $n \geq 2$. For $n = c-1$ this is also obvious, i.e. $\kappa_{\mathrm{XOR}}(c, c-1) = c/2$, and equation 2.2 implies that $\kappa_{\mathrm{XOR}}(c, c-3), \kappa_{\mathrm{XOR}}(c, c-2) \geq c/2$ and thus, by proposition 2.2, $= c/2$ as well; similarly $\kappa_{\mathrm{XOR}}(c-1, c) \geq c/2+1$ and thus $= c/2+1$. For $n = c/2$ we take the "odd coloring," namely I_T is $= 1, 0$ according to whether or not $|T|$ is odd. It is easy to see than any U has $|T \cap U|$ odd for at least half the T with $|T|$ odd, and so $\kappa_{\mathrm{XOR}}(c, c/2) \leq c/4+1$. Equation 2.1 implies that $\kappa_{\mathrm{XOR}}(c, c/2-1), \kappa_{\mathrm{XOR}}(c, c/2) \geq c/4+1$ and $\kappa_{\mathrm{XOR}}(c, c/2+1) \geq c/4+2$, and so by the above coloring and proposition 2.2 we are done.

□

For general n the problem of determining $\kappa_{\mathrm{XOR}}(c,n)$ is more difficult. However, for fixed c it suffices to check the cases $n = 1, 2, \cdots, O(c^2)$ to determine $\kappa_{\mathrm{XOR}}(c,n)$ for all n. That is, for a fixed $r \in [0, c-2]$ let $K = K(r)$ be the smallest integer such that for all m sufficiently large there exists a $(c; (c-1)m+r, (c/2-1)m+K+1)$ XOR coloring. By the above we have $K(r) \geq r\theta_c$, $K(0) = 0, K(1) = 1, K(c/2-1) = K(c/2) = c/4 = K(c/2+1)-1, K(c-3) = K(c-2) = c/2-2$. Proposition 2.2 implies $K(r) \leq c/2-2$ for all $r \in [2, c-4]$ and that for any $r \in [1, c-2]$, $K(r) - K(r-1)$ is either 0 or 1.

Lemma 2.4 *An XOR coloring with $I_T = 0$ for some T has $(k-1)/n \geq 1/2$. For any r there exists a unique $m_0 = m_0(r) \leq 2K(r) - r$ such that there exist $(c; (c-1)m + r, (c/2-1)m + K + 1)$ for all $m \geq m_0$.*

Proof The first statement follows from summing over all U with $|T \cap U| \equiv 1 \pmod 2$. For the second part, m_0 obviously exists, and the coloring at $n = (c-1)m_0 + r$ must have at least one I_T equal zero, or else we could subtract 1 from all the I_T's to get a coloring as above with $m = m_0 - 1$. So the first statement applies to yield $m_0 \leq 2K(r) - r$.

□

In particular, checking $\kappa_{\mathrm{XOR}}(c,n)$ for $n = 1, 2, \ldots, O(c^2)$, we can determine all $K(r)$'s, and therefore all $\kappa_{\mathrm{XOR}}(c,n)$ with $n \geq m_0(r)(c-1) + r = O(c^2)$ by the above. Another consequence of the above is:

Proposition 2.5 *$K(2) = 2$ for $c = 2^s \geq 8$, and for $n > 2$ and $\equiv 2 \pmod{c-1}$ we have $\kappa_{\mathrm{XOR}}(c,n) = 2 + \lceil n\theta_c \rceil$. In general $K(r) = \lceil r\theta_c \rceil$ iff for $n \equiv r \pmod{c-1}$ and n sufficiently large have $\kappa_{\mathrm{XOR}}(c,n) = 1 + \lceil n\theta_c \rceil$.*

Proof $K(1) = 1$ implies that either $K(2)$ is 1 or 2. The lemma implies that to rule out $K(2)$ being 1 it suffices to check the case $m_0 = 0$, i.e. $n = 2$; for $n = 2$ it is easy to see that given any $T_1, T_2 \subset S$ there exists a U with $|T_1 \cap U|$ and $|T_2 \cap U|$ even, provided that $s \geq 3$. Hence $K(2) = 2$, and $m_0(2) = 1$, i.e. for $n = c+1$ we can achieve $\kappa_{\mathrm{XOR}}(c, n) = 2 + \lceil n\theta_c \rceil$. The results on K clearly translate into the last sentence of the proposition.

□

Proposition 1.3 is a consequence of the above.

3 Almost Equicolored Colorings and Profiles

For a c-coloring, $\gamma: \mathbf{B}^n \to \mathbf{B}^s$, we define its $L^p(\mathcal{H}_k)$ distance from equicolor via

$$\|\gamma\|^p_{L^p(\mathcal{H}_k)} = \sum_{v \in \mathbf{B}_s} \|\gamma^{-1}(v)\|^p_{L^p(\mathcal{H}_k)},$$

where for a $C \subset \mathbf{B}^n$ (and a fixed $c = |C|/n$ in mind) we define the summand via

$$\|C\|^p_{L^p(\mathcal{H}_k)} = \sum_{H \in \mathcal{H}_k} \Big| |C \cap H| - |H|/c \Big|^p.$$

This is one sense in which we can measure how close to being equicolored a coloring is. In this section we study the case $p = 2$. There are other important metrics suggested by the applications, and we mention

$$\|\gamma\|^p_{\mathrm{RP}(\mathcal{H}_k)} = \max_{H \in \mathcal{H}_k, G \subset \mathbf{B}^s, |G| = 2^{s-1}} |\gamma^{-1}(G) - 2^{k-1}|\,;$$

this measures how well the bits γ extracts work as a random source to an RP algorithm, G being interpreted as the set of witnesses.

We study $\|C\|_{L^2(\mathcal{H}_k)}$ for a $C \subset \mathbf{B}^n$ via C's profile in the following sense:

Definition 3.1 *The* profile *of a $C \subset \mathbf{B}^n$ is the collection of numbers, p_i, defined via*

$$p_i \equiv \frac{1}{|C|} \sum_{v \in C} \frac{|C \cap \Gamma_i(v)|}{|\Gamma_i(v)|},$$

where $\Gamma_i(v)$ denotes the set of vertices of $\mathbf{B}^n$ of distance exactly i to v (in paricular, $|\Gamma_i(v)| = \binom{n}{i}$). For a coloring $\gamma: \mathbf{B}^n \to \mathbf{B}^s$ we define its profiles *to be those of the unicolored subsets of $\mathbf{B}^n$, i.e. sets of the form $\gamma^{-1}(v)$ for some $v \in \mathbf{B}^s$; if these profiles are indentical we refer to the* profile *of γ.*

Intuitively p_i measures that probability that in picking a random $v \in C$ and then random node of distance i to v the new node will also lie in C.

Consider, for a $C \subset \mathbf{B}^n$ with $|C| = n/c$,

$$\sigma_j = \|C\|^2_{L^2(\mathcal{H}_k)} = \sum_{H \in \mathcal{H}_j} (|C \cap H| - |H|/c)^2 = \sum_{H \in \mathcal{H}_j} |C \cap H|^2 - \sum_{H \in \mathcal{H}_j} (|H|/c)^2.$$

These σ_j's measure the L^2 distance to being equicolored, σ_j vanishing precisely when the coloring is $\mathcal{H}_j$ equicolored. Clearly the σ_j can be expressed in terms of the p_i, via

$$\sum_{H \in \mathcal{H}_j} |C \cap H|^2 = \sum_{u,v \in C} \binom{n - \rho(u,v)}{j - \rho(u,v)}$$

(where ρ denotes the distance in $\mathbf{B}^n$) and a short calculation yields:

Proposition 3.2

$$\sigma_j = \frac{2^n}{c} \left(-\frac{2^j}{c} \binom{n}{j} + \sum_{l=0}^{j} M_{j,l}\, p_l \right).$$

where

$$M_{j,l} = \binom{n-l}{j-l}\binom{n}{l} = \binom{n}{l, j-l, n-j} = \binom{n}{j}\binom{j}{l}.$$

Let A_i denote the i-th neighbor matrix of $\mathbf{B}^n$, i.e. with a 1 or a 0 in the v, w entry according to whether or not v, w are vertices of distance i in $\mathbf{B}^n$. We have

$$p_i = \frac{(A_i \chi_C, \chi_C)}{|C| \binom{n}{i}}.$$

We now give a lower bound for σ_k, which also proves theorem 1.6. Before doing so, we remark that if $c - 1 | n$, we expect that the balanced XOR yields the lowest σ_j. It is easy to see that the profile of any balanced XOR, which we denote $\tilde{p}_i$, is given by the recurrence,

$$\tilde{p}_{i+1} = \frac{n}{n-i} \frac{1 - \tilde{p}_i}{1 - c} - \frac{i}{n-i} \tilde{p}_{i-1} \quad \forall i \geq 2, \qquad \tilde{p}_0 = 1,\ \tilde{p}_1 = 0$$

(see the next section). In particular, all inequalities derived in this paper for σ_j are designed to be tight for the $\tilde{p}_i$'s.

We notice that the p_i can be related to the μ_r's of last section. Namely, it is easy to see that

$$A_i A = (i+1) A_{i+1} + (n - i + 1) A_{i-1},$$

and hence we can write the $A_i = q_i(A)$, where q_i is a polynomial of degree i given by

$$q_{i+1}(x) = \frac{xq_i(x)}{i+1} - \frac{n-i+1}{i+1}q_{i-1}(x), \quad \forall i \geq 2, \qquad q_0 = 1,\ q_1 = x.$$

So consider the polynomials

$$s_j(x) = \sum_{l=0}^{j} \frac{\binom{j}{l}}{\binom{n}{l}} q_l(x). \tag{3.1}$$

By proposition 3.2 we have

$$\sigma_j = \frac{2^n}{c}\binom{n}{j}\left(-\frac{2^j}{c} + \frac{(s_j(A)\chi_C, \chi_C)}{|C|}\right).$$

Proposition 3.3 *The s_j's are given by*

$$s_{j+1}(x) = \frac{x + (n-2j)}{n-j} s_j(x) \quad \forall j \geq 1, \qquad s_0(x) = 1.$$

Proof It is clear from the definition of the s_j's that they are polynomials of degree j satisfying

$$(s_j(A)f, f) = \sum_{H \in \mathcal{H}_j} \left(\sum_{x \in H} f(x)\right)^2,$$

for any $f \in L^2(\mathbf{B}^n)$. The right-hand-side of the above clearly vanishes if f is the XOR of r variables with $r \geq n - j + 1$. Since such an f has eigenvalue $n - 2r$, we conclude that s_j has roots $-n, -n+2, \ldots, -n+2(j-1)$. It remains to determine the leading coefficient; clearly q_i has leading coefficient $1/i!$, and from equation 3.1 it follows that that of s_j is $1/\binom{n}{j}$ times that of q_j.

□

From this we can give a general lower bound on σ_j:

Theorem 3.4 *For any $j < 1 + \theta_c n$ we have*

$$\frac{\left(s_j(A)\chi_C, \chi_C\right)}{(\chi_C, \chi_C)} \geq \frac{1}{c}s_j(n) + \frac{c-1}{c}s_j\big(-n/(c-1)\big),$$

equality holding iff $c-1|n$ and all μ_i's vanish except for $\mu_0 = 1/c$ and $\mu_{n(1-\theta_c)} = (c-1)/c$. If $-n/(c-1) = t + \alpha$ with t an integer and $\alpha \in (0,1)$, we also have the sharper bound

$$\frac{\left(s_j(A)\chi_C, \chi_C\right)}{(\chi_C, \chi_C)} \geq \frac{1}{c}s_j(n) + \frac{c-1}{c}\big((1-\alpha)s_j(t) + \alpha s_j(t+1)\big),$$

with equality iff $\mu_t = (1-\alpha)(c-1)/c$, $\mu_{t+1} = \alpha(c-1)/c$. In particular, for $c-1|n$ any optimal coloring for σ_j has the same profile as that of the balanced XOR coloring.

Proof The function

$$g(x) \equiv \begin{cases} s_j(x) & \text{if } x \geq 2(j-1)-n \\ 0 & \text{otherwise} \end{cases}$$

attains the same values as $s_j(x)$ on the eigenvalues of A, and so

$$\frac{\left(s_j(A)\chi_C, \chi_C\right)}{(\chi_C, \chi_C)} = \sum_{r=0}^{n} s_j(n-2r)\mu_r = \sum_{r=0}^{n} g(n-2r)\mu_r$$

Recall that $\mu_r \in [0,1]$, $\mu_0 = 1/c$, and since $(A\chi_C, \chi_C) \geq 0$ we have

$$\sum_{r=0}^{n}(n-2r)\mu_r \geq 0;$$

this implies that

$$\sum_{r=1}^{n}\mu_r = \frac{c-1}{c}, \qquad \sum_{r=1}^{n}(n-2r)\frac{\mu_r}{\sum_{m=1}^{n}\mu_m} \geq \frac{-n}{c-1}.$$

But $g(x)$ is convex, and is strictly convex and monotone increasing for $x \geq 2(j-1)-n$, and $2(j-1)-n \leq (2\theta-1)n = -n/(c-1)$. Hence the theorem follows from standard facts on convex functions.

4 More on the Profile

In this section we further study the profile of a $C \subset \mathbf{B}^n$ with $|C| = n/c$. We believe that understanding the profile is a more robust way of studying the problem of approximately equicolored colorings for different metrics. For example, for p an even integer we can define "higher-order" profiles which determine the L^p distance of C to $1/c$ dense; namely, for each $\binom{p-1}{2}$-tuple of integers $\{i_{\alpha,\beta}\}$ ranging over $1 \leq \alpha < \beta \leq p$, we count the number of p-tuples of points $(u_1, \ldots, u_p)$ with $\rho(u_\alpha, u_\beta) = i_{\alpha,\beta}$. So understanding such higher-order profiles might lead to understanding approximately equicolored colorings for the L^p metrics, $p \geq 2$. However, at present we cannot even give a full proof of theorems 1.1 or 1.6 via a direct analysis of the profile. Yet the profile does have certain interesting properties, which we comment on here.

We start by deriving some inequalities satisfied by the p_i. If $\rho(u,v) = i$ and $\rho(v,w) = j$, then $\rho(u,w) = i+j-2r$ for some r, and counting for a fixed u, w with $\rho(u,w) = i+j-2r$ the number of v satisfying the above, it follows that

$$A_iA_j = \sum_{r=0}^{\min(i,j)} A_{i+j-2r}\binom{i+j-2r}{i-r}\binom{n-(i+j-2r)}{r}.$$

It is tempting to try to derive inequalities based on the above. Let us write $f_i = A_i\chi_C$. Notice that for C given by a balanced coloring, the f_i and f_j won't be proportional, and applying Cauchy-Schwartz directly to (f_i, f_j) will yield inequalities which are not all tight in this case; what is true is that f_i and f_j are constant on C and on its complement, $\overline{C}$, and so we make the following definition.

Definition 4.1 *We say that f_i and f_j are* positively (evenly, negatively) alligned *if*

$$\frac{1}{|C|}(f_i, f_j) \; > \; \binom{n}{i}\binom{n}{j}\left(p_ip_j + \frac{(1-p_i)(1-p_j)}{c-1}\right)$$

(respectively, $=$ and $<$).

Proposition 4.2 *For any i, j we have*

$$\sum_{r=0}^{\min(i,j)} p_{i+j-2r}\binom{n}{i-r, j-r, r, n-i-j+r}\left\{\begin{array}{c}>\\=\\<\end{array}\right\}\binom{n}{i}\binom{n}{j}\left(p_ip_j + \frac{(1-p_i)(1-p_j)}{c-1}\right),$$

according to the allignment of f_i and f_j.

Clearly f_i and f_j are evenly alligned if both are constant on C and $\overline{C}$, and so for the balanced coloring profile we obtain

$$\tilde{p}_{j+1} = \frac{n}{n-j}\left[\tilde{p}_1\tilde{p}_j + \frac{(1-\tilde{p}_1)(1-\tilde{p}_j)}{c-1}\right] - \frac{j}{n-j}\tilde{p}_{j-1}, \quad \forall j \geq 1.$$

Proposition 4.3 *f_i is nonnegatively alligned with itself. In particular, proposition 4.2 holds with $\geq$ for $i = j$.*

Proof We have

$$(f_i, f_i) = (f_i, f_i)_C + (f_i, f_i)_{\overline{C}},$$

where $(\cdot\,,\,\cdot)_C$ denotes the inner product of $\mathbf{B}_n$ functions restricted to C (i.e. summing over only vertices of C) and similarly for $(\cdot\,,\,\cdot)_{\overline{C}}$. For $v \in C$, the average value of f is $p_i\binom{n}{i}$; for $v \notin C$, the average value is $(1-p_i)\binom{n}{i}/(c-1)$. Applying Cauchy-Schwartz we have

$$(f_i, f_i)_C \geq \frac{(f_i, 1)_C^2}{(1,1)_C},$$

where 1 denotes the all 1's vector, and similarly for $\overline{C}$, from which the proposition follows.

□

The following is another approach to the approximate equicoloring problem. If it were true that f_i were always nonnegatively correlated with f_1, then it would follow that

$$p_{j+1} \geq \frac{n}{n-j}\left[p_1 p_j + \frac{(1-p_1)(1-p_j)}{c-1}\right] - \frac{j}{n-j}p_{j-1}.$$

If so, one could inductively bound the partial sums

$$\tau_{j,l} = \sum_{i=l}^{j} \binom{j}{i} p_i,$$

as

$$\tau_{j,l} \geq \alpha_l p_l + \beta_l p_{l-1} + \gamma_l,$$

for constants $\alpha_l, \beta_l, \gamma_l$ depending only on p_1. The inductive step uses the inequality on p_l in terms of p_{l-1} and p_{l-2}, and the induction continues as long as α_l is positive; the inductive step yields

$$\alpha_{l-1} = \frac{j-l+1}{m}\left(\beta_l + \frac{n}{n-l}\alpha_l \eta\right), \quad \beta_{l-1} = 1 - \frac{j-l+1}{m}\left(\alpha_l \frac{l}{n-l}\right),$$

with $\alpha_j = 1, \beta_j = 0$, where $\eta = p_1 - (1-p_1)/(c-1)$ (the γ_l are unimportant in checking the positivity of the α_l's). Numerical experiments done by the author indicate that α_l always remain positive. If generally true, this would give another proof of theorem 1.6, in cases where f_l and f_1 are always nonnegatively correlated.

Unfortunately, it can happen that f_l and f_1 are negatively alligned. Namely, if C lies in the subsets of vertices with, say, $x_1 + \cdots + x_n = 1$, then $(f_l, f_1) = 0$ for all l even. It would be interesting to see if, say, in such a case there would be enough positive allignment between f_1 and f_l for l odd so as to allow a modification of this method to deduce theorems 1.1 and 1.6.

We finish this section with the following curious observation about classifying the profiles which give $\sigma_k = 0$ for a fixed k. Such p_i's must satisfy $\sigma_j = 0$ for all $j \geq k$. So consider the system of equations in the variables $(p_0, \ldots, p_n; \nu)$:

$$\sum_{l=0}^{j} \binom{n-l}{j-l}\binom{n}{l} p_l = \nu \frac{2^j}{c}, \quad \forall j,\ k \leq j \leq n, \qquad p_0 = \nu. \tag{4.1}$$

In the above we have thrown in a dummy parameter ν to make the system homogeneous; in our case, $\nu = p_0 = 1$. Although the p_i satisfy further constraints, this linear system alone is quite interesting. One can write a very simple basis for the solutions $(p_0, \ldots, p_n; \nu)$ of this system which has some curious properties.

For any integer r let

$$P^r = \left(0, 1, -2^r, 3^r, \ldots, (-1)^{n+1} n^r; 0\right),$$

and let $E = (1, 1, \ldots, 1; 1)$.

Proposition 4.4 *The collection $E, P^1, P^2, \ldots, P^{k-1}$ form a basis of the solution set of equation 4.1. In other words, solutions are characterized by $\nu = p_0$ and the recurrence $(\sigma+1)^{k-1}(\sigma-1)p_i = 0$ for $i \geq k$, where σ is the index downshifting operator, $\sigma(p_t) = p_{t-1}$.*

The above vectors are clearly linearly independent, and it is easy to check that E satisfies equation 4.1. To verify that the above P^r are also solutions, it suffices to check that $P^{(r)}$, whose general p_i term is $(-1)^{n+1}n^{(r)}$ where $n^{(r)} = n(n-1)\cdots(n-r+1)$, satisfies equation 4.1; this easily reduces to the fact that the sum of alternating binomial coefficients, $\binom{r}{0} - \binom{r}{1} + \binom{r}{2} - \cdots$ is zero. The first part of equation 4.1 is a system of $n-k+1$ equations lower triangular in $p_k, \ldots, p_n$; these equations and the last, $p_0 = \nu$ are therefore linearly independent.

□

An interesting easy consequence is

Corollary 4.5 *The $\tilde{p}_i$'s are given by $\tilde{p}_j = (-1)^j r(j) - 1/c$ for some polynomial r of degree $k-1 = n\theta_c$.*

We also mention one more solution to equation 4.1, which is rather curious. Let $\hat{P} = (\hat{p}_0, \ldots, \hat{p}_n; 1)$ be the vector with $\hat{p}_i$ determined by the conditions

$$\hat{p}_0 = 1, \qquad \hat{p}_1 = \cdots = \hat{p}_{k-1} = 0,$$

and $\hat{p}_j$ for $j \geq k$ recursively determined via

$$\hat{p}_j + \binom{k}{1}\hat{p}_{j-1} + \cdots + \binom{k}{k}\hat{p}_{j-k} = 2^k/c. \tag{4.2}$$

Proposition 4.6 *$\hat{P}$ is a solution to equation 4.1. Assume $2^k/c \geq 4$. For $i \geq k$ the $\hat{p}_i$'s are integers, alternating in sign, and increasing in absolute value. equation 4.1.*

Proof $\hat{P}$ clearly satisfies $(\sigma+1)^{k-1}(\sigma-1)\hat{p}_i = 0$ for $i \geq k$. So we have $\hat{p}_j = (-1)^j q(j) - 1/c$ for some polynomial q of degree $\leq k-1$. Assuming $2^k/c \geq 4$, since $\hat{p}_0 = 1$ and $\hat{p}_1 = \cdots = \hat{p}_{k-1} = 0$, it is easy to see that q has one root in each of $(-\infty, 0), (1,2), (2,3), \ldots, (k-2, k-1)$, and then that the absolute values of the $\hat{p}_i$ is strict increasing.

5 Locally Symmetric Colorings and Concluding Remarks

We consider the problem of classifying for $c-1|n$ all optimal colorings, i.e. with $k = 1 + n\theta_c$. We call a coloring, $\gamma: \mathbf{B}^n \to \mathbf{B}^s$ *locally symmetric* if there is an η such

that for all $v \in \mathbf{B}_n$, exactly $n\eta$ of v's neighbors are colored $\gamma(v)$, and every other color occurs among v's neighbors exactly $n(1-\eta)/(c-1)$ times. In addition we say that the coloring is *sparse* if $\eta = 0$. The balanced coloring has this property.

More generally we define a coloring to be Γ_k-symmetric if there is an η such that for every $v \in \mathbf{B}^n$ exactly $\binom{n}{k}\eta$ of $\Gamma_k(v)$ have the same color as v, and the other colors each occur $\binom{n}{k}(1-\eta)/(c-1)$ times. An easy induction argument shows that any locally symmetric coloring is also Γ_k-symmetric for all k, and that their profile is determined by $p_0 = 1$, $p_1 = \eta$, and

$$p_{j+1} = \frac{n}{n-j}\left[p_1 p_j + \frac{(1-p_1)(1-p_j)}{c-1}\right] - \frac{j}{n-j}p_{j-1}, \quad \forall j \geq 1.$$

Since the p_i's are determined by the μ_r's, an optimal locally symmetric coloring for $c-1|n$ must have $\eta = 0$, i.e. must be sparse.

We also remark that locally symmetric colorings are precisely those whose unicolor sets, C, have $f_i = A_i\chi_C$ constant on C and $\overline{C}$. Such colorings have f_i, f_j evenly alligned for all i, j.

We pose two questions, which we cannot resolve at this point:

Problem 5.1 *Are all optimal colorings necessarily locally symmetric, at least if* $c-1|n$*?*

Problem 5.2 *Are all locally symmetric sparse colorings necessarily XOR colorings?*

Regarding the latter question, and XOR colorings in general, one can make the following observation. Consider a general cycle of length 4 in $\mathbf{B}_n$, $(v_0, v_1, v_2, v_3, v_0)$. XOR colorings satisfy the following two conditions:

1. $\gamma(v_3)$ depends only on $\gamma(v_0), \gamma(v_1), \gamma(v_2)$, not on the particular cycle,
2. $\gamma(v_0) = \gamma(v_2)$ implies $\gamma(v_1) = \gamma(v_3)$.

Proposition 5.3 *Any locally symmetric coloring satisfying the above two conditions is an XOR coloring.*

Proof Fix any coloring, γ, of $\mathbf{B}^n$, and a vertex, $v \in \mathbf{B}_n$. We define a group law on the colors as follows. We can assume $\eta < 1$, or else there is nothing to prove. So for any $\gamma_1 \neq \gamma_2$ neither equal $\gamma(v)$, there exist neighbors v_1, v_2 of v with $\gamma(v_i) = \gamma_i$. There exists a unique v_3 making v, v_1, v_3, v_2 a simple cycle of length 4; define $\gamma_1 + \gamma_2$ to be $\gamma(v_3)$. If $\gamma_1 = \gamma_2$ define their sum to be $\gamma(v)$, and if one of γ_1, γ_2 is $\gamma(v)$ define their sum to be the other γ_i. That this defines defines a commuatative, associative group law, with identity $\gamma(v)$, and every other element of order 2, is an easy consequence of the above conditions; for example, commutativity follows from the fact that if v_0, v_1, v_2, v_3 is a simple cycle then so is v_0, v_3, v_2, v_1; associativity follows from the fact that the

sum of three colors is the antipodal point to v_0 in a subcube of $\mathbf{B}^n$ isomorphic to $\mathbf{B}^3$. This sets up an isomorphism between the colors and $\mathbf{B}^s$. One can similarly show that for every cycle $(v_0, v_1, v_2, v_3, v_0)$ we have the sum of the colors vanishing (i.e. $= \gamma(v)$). Setting coordinates on $\mathbf{B}^n$ so that v is the origin, the coloring of the neighbors of v determine XOR's, $f_1, \ldots, f_s$ such that the f_i's induce the given coloring on v and its neighbors. The conditions on the coloring imply, by induction on k, that for any k the coloring determined by the $f_1, \ldots, f_s$ agrees with the original coloring on $\Gamma_k(v)$.

□

Of course, in the above proposition, one can replace the local symmetry condition by the condition that, say, each $v \in \mathbf{B}^n$ has all colors appearing among its neighbors except for, possibly, its own color. Call such a coloring *connected.* We can restate the first problem as:

Problem 5.4 *Is any optimal coloring necessarily connected and does it necessarily satisfy the above two conditions?*

References

[BBR88] C.H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM J. on Computing*, 17(2):210–229, 1988.

[Bra89] G. Brassard. Cryptology— Column 1. *SIGACT News*, 20(3):15–19, Summer 1989.

[CFG+85] B. Chor, J. Friedman, O. Goldreich, J. Håstad, S. Rudich, and R. Smolensky. The bit extraction problem or t-Resilient functions. In *26th Annual Symposium on Foundations of Computer Science*, pages 396–407, 1985.