On the Number of Monochromatic Close Pairs of Beads in a Rosary

September 1984
(revised May 1986)

Oded Goldreich

MIT – Laboratory for Computer Science
Cambridge, Massachusetts 02139

Abstract — We consider the following problem: Let r be a n-bead rosary with m white beads and $n-m$ black beads. Let t be an integer, $t \ll n$. Denote by $MC_t(r)$ the number of pairs, of monochromatic beads which are within distance t apart, in the rosary r. What is the minimum value of $MC_t(\cdot)$, when the minimum is taken over all n-bead rosaries which consists of m white beads and $n-m$ black beads?
We prove a (reasonably) tight lower bound for this combinatorial problem. Surprisingly, when $m = n/2$, the answer is $\approx (\sqrt{2}-1) \cdot nt$, rather than $nt/2$ that one might have expected.

1. INTRODUCTION

This article addresses the following problem: For integers n, m, t ($n/2 \le m < n$, $t \ll n$) consider cyclic strings of m ones and $n-m$ zeros. Count the number of pairs of equal bits which are at most t places apart. What is the minimum of this count?

As one might have expected, the answer is essentially* $v_t^\rho \cdot nt$, where v_t^ρ is a constant depending only on t and $\rho \overset{\text{def}}{=} m/n$ (the fraction of ones in the string). However, the expression we get for v_t^ρ is somewhat surprising:

$$v_t^\rho = \sqrt{(4 + 8\rho^2 - 8\rho) \cdot \frac{t+1}{t}} - \frac{t+1}{t}$$

In particular, $v_t^{1/2}$ converges to $\sqrt{2}-1 \approx 0.414$ (not to $1/2$!), when t grows.

The above combinatorial problem occured to us when trying to analyze the performence of a special purpose oracle-sampling technique (for more details see our technical report [1]). An alternative formulation of the problem was suggested by one of the referees. Let n, m, t be integers as above. Let $G_{n,t}$ be the graph with vertex set $\{1, 2, ..., n\}$, where i and j are adjacent if $|i-j| \le t$ or $n - |i-j| \le t$. What is the largest cut in $G_{n,t}$ with m vertices on one side and $n-m$ vertices on the other side?

---

* Ignoring additive "error" terms of the form $O(t^2 + n/t)$.

2. DEFINITIONS AND CONVENTIONS

Let $s = (s_0, s_1, s_2, .., s_{n-1})$ be a binary string of length $n \stackrel{\rm def}{=} |s|$. Following the description of the introduction, we let $c_t(s)$ count the number of equal and close bits. Namely

$$c_t(s) \stackrel{\rm def}{=} |\{(i,j):\ 0 \le i < j < n \ \wedge\ s_i = s_j \ \wedge\ \delta(i,j) \le t\}| \,,$$

where $\delta(i,j)$ is the cyclic distance between i and j (i.e. $\delta(i,j) = \min\{|j-i|, n-|j-i|\}$). An alternative definition of c_t follows (indices are computed modulo n)

$$c_t(s) = \sum_{i=1}^{t} |\{j:\ 0 \le j < n \ \wedge\ s_j = s_{j+i}\}| \,.$$

Let n and m be integers such that $0.5n \le m < n$. Let $\rho \stackrel{\rm def}{=} \frac{m}{n}$. We denote by S_n^ρ the set of n-bit binary strings with $m = \rho n$ ones (and $n - m$ zeros). Denote by $C(n, \rho, t)$ the minimum value of the count $c_t(\cdot)$ divided by nt, when minimized over all strings in S_n^ρ. That is

$$C(n, \rho, t) = \frac{1}{nt} \cdot min_{s \in S_n^\rho} \{c_t(s)\} \,.$$

Throughout the article, we assume that $t < n/2$ and $t > \rho/(1-\rho)$. The other cases are less interesting and easily reducible to the case we consider. Further details can be found in our technical report [1].

Proposition 1: Let $sh_i(s) = (s_i, s_{i+1}, s_{i+2}, ..., s_{i+n-1})$. Then $c_t(s) = c_t(sh_i(s))$.

Prop. 1 follows directly from the definitions which consider strings as if they were cycles. From this point on, we also take the liberty of doing so.

3. LOWER BOUND ON $C(n, \rho, t)$

We will analyze $C(n, \rho, t)$ as follows: first we will show that the minimum of $c_t(\cdot)$ is acheived by strings which belong to a restricted subset of S_n^ρ; and next we will minimize $c_t(\cdot)$ over this subset. This will establish a lower bound on $C(n, \rho, t)$.

When evaluating $c_t(s)$, it may be of use to consider "lines" which connect positions that contain equal values and are less than t bits apart in the string s. Since $t < \frac{n}{2}$, there is only one way to draw the lines. These lines are hereafter called *overlines*. Note that $c_t(s)$ is nothing but the number of overlines in the string s.

3.1 Reduction into a restricted subset

In this subsection we will show that when analysing $C(n, \rho, t)$ it suffices to consider strings in S_n^ρ which have the following two properties:

[a] The string contains *no* short 3-alternating substrings (see Definition 1 below).

[b] The string contains *no* long homogenous substrings (see Definition 2 below).

Definition 1: A *3-alternating substring* is a substring of the form $\sigma^+\tau^+\sigma^+\tau^+$, where $\sigma \neq \tau \in \{0,1\}$. (Here, and throughout this article, σ^+ denotes a non-empty string of σ's.)
A 3-alternating substring is called *short* if it has length at most $t+1$.

Definition 2: A *long homogenous substring* is a substring of the form σ^{t+1}, where $\sigma \in \{0,1\}$.

We first build up tools to prove that it suffices to consider strings with no short 3-alternating substrings (Prop. 2 through 6, culminating in Lemma 1). Next we prove that with no loss of generailty, also the second condition holds (Lemma 2).

3.1.1 Getting rid of short 3-alternating substrings

Proposition 2: Let $\sigma_j \in \{0,1\}$, for $1 \leq j \leq 2t$. Let α be an arbitrary binary string. Then $c_t(\sigma_1\sigma_2\cdots\sigma_t 10\sigma_{t+1}\sigma_{t+2}\cdots\sigma_{2t}\alpha) - c_t(\sigma_1\sigma_2\cdots\sigma_t 01\sigma_{t+1}\sigma_{t+2}\cdots\sigma_{2t}\alpha) = 2\cdot(\sigma_1 - \sigma_{2t})$.
proof: The difference between the two counts is only due to the existence or non-existence of overlines between σ_1 and 1 and between 0 and σ_{2n} . Details are left to the reader. □

Note that *switching* τ_1 and τ_2 in the string $\sigma_1\sigma_2\cdots\sigma_t\tau_1\tau_2\sigma_{t+1}\sigma_{t+2}\cdots\sigma_{2t}\alpha$ results in the string $\sigma_1\sigma_2\cdots\sigma_t\tau_2\tau_1\sigma_{t+1}\sigma_{t+2}\cdots\sigma_{2t}\alpha$. The latter string has more overlines (than the former one) only in the case that $\sigma_1 = \tau_2 \neq \tau_1 = \sigma_{2n}$.

Proposition 3: Let α be a binary string, $\sigma \neq \tau \in \{0,1\}$ and let x, y, z, u be integers such that $x + y \geq t$ but $y + z < t$. Then:
[a] $c_t(\sigma\tau^x\sigma^y\tau^{z-1}\sigma\tau\alpha) \leq c_t(\sigma\tau^x\sigma^y\tau^z\sigma\alpha)$.
[b] $c_t(\sigma\tau^x\sigma^y\sigma\tau^z\alpha) \leq c_t(\sigma\tau^x\sigma^y\tau^z\sigma\alpha)$.
proof: Part (a) follows by switching in $\sigma\tau^x\sigma^y\tau^z\sigma\alpha$ the σ on the l.h.s. of α with the τ on the l.h.s. of that σ; and recalling Prop. 2. (Notice that the symbol in $\sigma\tau^x\sigma^y\tau^z\sigma\alpha$ which is t bits to the left of "the switched τ" is also a τ.) Part (b) follows by z sequential applications of part (a). □

Prop. $3_{(b)}$ will be used in order to get rid of short 3-alternating substrings. This will be done by scanning the string from left to right. Suppose that the string has the form $\alpha_1\tau^x\sigma^y\tau^z\sigma\alpha_2$, where the $\alpha_1\tau^x\sigma^y$ part contains no short 3-alternating substrings and $y + z < t$ (i.e. $\tau\sigma^y\tau^z\sigma$ is a short 3-alternating substring). Applying Prop. $3_{(b)}$, we transform the string to $\alpha_1\tau^x\sigma^{y+1}\tau^z\alpha_2$ (without increasing the number of overlines). This is repeated untill the $\sigma^+\tau^+$ substring following $\alpha_1\tau^x$ has length greater or equal to t.
Minor but crucial details which need to be considered are:

[1] The procedure is initiated with α_1 being the empty string. But how is one guaranteed to have a substring of the form $\tau^x\sigma^y$ with $x + y \geq t$? The answer is given by Prop. 4.

[2] The procedure is terminated when α_2 is empty. At this point there may be two short 3-alternating substrings. A better analysis shows that there may be only one (see Prop. 5). Finally, we get rid of the possibly remaining short 3-alternating substring (see Prop. 6).

Proposition 4: Let $s \in S_n^\rho$ be a binary string such that $c_t(s) = nt \cdot C(n,\rho,t)$. Then there exist a string, $s' \in S_n^\rho$, such that both the following conditions hold:

[a] The string s' contains **a** substring of the form 10^+1^+0 the length of which is **at least** $t+2$.

[b] $c_t(s') < c_t(s) + t^2$.

proof: W.l.o.g., s is not of the form 0^+1^+. Consider an arbitrary substring, α, of length t in s. Let z denote the number of zeros in α. Replacing α by 0^z1^{t-z} in the string s results in a string s', which satisfies condition (a). It is easy to see that $c_t(s') \leq c_t(s) + t(t-1)$. □

Proposition 5: Let $s' \in S_n^\rho$ be a string, with the *minimum* number of overlines, which satisfies Prop. 4. (Recall that $c_t(s') < nt \cdot C(n,\rho,t) + t^2$.) Then with no loss of generality, the string s' contains *at most one* short 3-alternating substring.

proof's sketch: By the hypothesis, s' contains a substring of length at least $t+2$ which has the form 10^+1^+0. Using the procedure outlined above (after Prop. 3), we scan s' and transform it so that none of the *scanned* 3-alternating substrings is short. We stop before scanning the last unscanned 01^+0^+1 substring. The reader may easily verify that the above process does not increase the number of overlines, since Prop. $3_{(b)}$ is used in the substitutions. For more details, see [1]. □

Proposition 6: Let $s' \in S_n^\rho$ be a string as in Prop. 5. Then there exist a string $s'' \in S_n^\rho$ satisfying the following two conditions:

[a] The string s'' contains no short 3-alternating substring.

[b] $c_t(s'') < c_t(s') + t^2$.

proof: By the hypothesis s' contains at most one short 3-alternating substring. Assume that such a unique 01^y0^z1 substring of length less than $t+2$ does exist (i.e. $y+z<t$). Replacing this substring in s' by the substring 00^z1^y1 results in a string s''. Note that s'' satisfies (a). To conclude note that $c_t(s'') < c_t(s') + t^2 - t$. The proposition follows. □

Definition: Let R_n^ρ be the set of strings which belong to S_n^ρ and do not have short 3-alternating substrings. $C_R(n,\rho,t)$ will denote $\min_{r \in R_n^\rho} \frac{1}{nt} \cdot c_t(r)$.

Lemma 1: $C(n,\rho,t) > C_R(n,\rho,t) - \frac{2t}{n}$.

proof: Immediate by Prop. 4, 5 and 6. □

3.1.2 Getting rid of long homogenous substrings

We now define even a more restricted subset of S_n^ρ:

Definition: The set MR_n^ρ is the subset of strings which belong to R_n^ρ and do not have long homogenous substrings. $C_{MR}(n,\rho,t)$ will denote $\min_{r \in MR_n^\rho} \frac{1}{nt} \cdot c_t(r)$.

Next, we show that a string, $r_0 \in R_n^\rho$, with minimum overlines can be transformed into a string $r_0' \in MR_{n'}^{\rho'}$, such that $n' \approx n$, $\rho' \approx \rho$ and $c_t(r_0') \approx c_t(r_0)$.

Proposition 7: Let $r_0 \in R_n^\rho$ be a string with minimum number of overlines (i.e. $c_t(r_0) = nt \cdot C_R(n,\rho,t)$). Then:

[a] For $\sigma \in \{0,1\}$, if r_0 contains a substring of more than t consecutive σ's then r_0 contains no block of less than t consecutive σ's. Futhermore, without loss of generality, r_0 contains at most one substring of more than t consecutive σ's.

[b] The string r_0 has no substring of the form σ^{2t}.

[c] There exist a $k < t$, a $\rho' \geq \rho$ and a $r_0' \in MR_{n+k}^{\rho'}$ such that $c_t(r_0) \geq c_t(r_0') - kt$.

proof:

Part (a): Omitting one σ from a substring that contains more than t σ's decreases the number of overlines by exactly t. Adding one σ to a block of k σ's increases the number of overlines by t if $k \geq t$, and by less than t if $k < t$. Part (a) of the proposition follows easily.

Part (b): Assume on the contrary that r_0 contains a σ^{2t} substring, and let $\tau \neq \sigma \in \{0,1\}$. We first note that in both cases ($\sigma \in \{0,1\}$), the string r_0 contains a $\tau\tau$ substring. We omit a single τ from the $\tau\tau$ substring and insert it in the middle of the $\sigma^t\sigma^t$ substring, decreasing the number of overlines and yielding a contradiction.

Part (c): By part (a), r_0 contain at most one 0^t0^+ [1^t1^+] block. Also, if r_0 contains a 0^{t+j} substring then it contains also a 1^{t+j} substring. Let l denote the length of the longest 1^+ substring in r_0. By part (b), $l < 2t$. In case $l \leq t$, we are done. The interesting case is when $t < l < 2t$. Set $k = 2t - l$ and r_0' to be the string which results from r_0 by the following procedure:

step 1: add k ones to the longest 1^+ block (yielding a 1^{2t} block);

step 2: if r_0 contains a 0^{t+u} block (when $u > 0$) then omit u zeros from the 0^{t+u} block and insert them in the middle of the 1^{2t} block.

step 3 (Recall that r_0 contains a 00 substring): if r_0 does not contain a 0^{t+1} block then omit a single 0 from a 00 substring and insert it in the middle of the 1^{2t} block.

Note that $\rho' = \frac{\rho n+k}{n+k}$ is the fraction of ones in r_0' (i.e. $r_0' \in MR_{n+k}^{\rho'}$). It is easy to see that $c_t(r_0') < c_t(r_0) + kt$ and that $\rho' = \rho + \frac{(1-\rho)k}{n+k} > \rho$. Part (c) of the proposition follows. $\square$

Proposition 8: There exist $0 \leq k < t$ and $\rho' \geq \rho$ such that

$$C_R(n,\rho,t) > C_{MR}(n+k,\rho',t) - \frac{t}{n}\,.$$

proof: By Prop. $7_{(c)}$, $C_R(n,\rho,t) = \frac{1}{nt}\cdot c_t(r_0) > \frac{1}{nt}\cdot(c_t(r_0') - t^2) \geq C_{MR}(n+k,\rho',t) - \frac{t}{n}$. $\square$

Lemma 2: Let $v(\rho,t)$ be a function which increases monotonely with ρ (when $\rho \geq 1/2$).

$$\text{If } C_{MR}(n,\rho,t) \geq v(\rho,t) \text{ then } C_R(n,\rho,t) \geq v(\rho,t) - \frac{t}{n}\,.$$

proof: Immediate by Prop. 8. $\square$

3.2 Lower bound for $C_{MR}(n,\rho,t)$

Recall that each of the strings in $MR_n^\rho \subset S_n^\rho$ has the following properties:

[a] The string contains no short 3-alternating substrings.

[b] The string contains no long homogenous substrings.

3.2.1 Introducing localized counting

We will relay on the above properties of the strings in MR_n^ρ in order to bound $C_{MR}(n,\rho,t)$. Given a string $r \in MR_n^\rho$ we will introduce an expression, for $c_t(r)$, which depends only on the numbers of bits in each maximal substring of consecutive equal bits. In other words, we will introduce a localized counting of $c_t(r)$.

Definition: We say that b is a *block* (an *all-σ-block*) of the string r if it is a maximal substring of equal bits. That is $b = \sigma^+$ and $r = \tau b \tau \alpha$, where $\tau \neq \sigma$ and α is an arbitrary string.

Notations: Let q denote the number of all-zero [all-one] blocks in r. Beginning from an arbitrary position between an all-one block and an all-zero block and going cyclically from left to right; number the blocks of consecutive zeros [ones] by 0,1,2,...,$(q-1)$. Denote by z_i the number of zeros in the i-th all-zero-block and by y_i the number of ones in the i-th all-one-block. That is, $r = 0^{z_0}1^{y_0}0^{z_1}1^{y_1}0^{z_2}1^{y_2}\cdots 0^{z_{q-1}}1^{y_{q-1}}$.

Proposition 9: Let $r \in MR_n^\rho$. Overlines occur (in r) only **either** within a block **or** between two consecutive blocks (of the same bit).

proof: Immediate from the fact that r does not contain short 3-alternating substrings. $\square$

The above suggests evaluating the number of overlines (in r) by counting the "contribution" of each block to it. This counting proceeds as follows:

Block-Localized Counting (with respect to a block of length l in r):

[a] The number of overlines **within the block**, denoted I_l.

[b] The number of overlines **between** bits of the **blocks** neighbouring this block (i.e the first block on its left and the first block on its right), denoted B_l.

Notations: Let $f(l)$ denote the total "contribution" of a l-bit long block. That is

$$f(l) \stackrel{\text{def}}{=} I_l + B_l$$

Proposition 10: Let $r \in MR_n^\rho$.

[a] $c_t(r) = \sum_{i=0}^{q-1} f(y_i) + \sum_{i=0}^{q-1} f(z_i)$, where $r = 0^{z_0}1^{y_0}0^{z_1}1^{y_1}\cdots 0^{z_{q-1}}1^{y_{q-1}}$.

[b] For $l < t$, $I_l = \binom{l}{2}$ and $B_l = \sum_{i=1}^{t-l} i$. For $l = t$, $I_l = \binom{t}{2}$ and $B_l = 0$.

[c] If $1 \le l \le t$ then $f(l) = l^2 - (t+1)l + \frac{t^2+t}{2}$.

proof: Part (a) follows by Prop. 9. One can easily verify the validity of Parts (b). Part (c) follows immediately from Part (b). $\square$

3.2.2 Finding the minimum

Notations: Let

$$g(x_0, x_1, .., x_{q-1}) \stackrel{\text{def}}{=} \sum_{i=0}^{q-1} f(x_i)$$

Proposition 11: For fixed q, t and k, the minimum value of the function $g(x_0, .., x_{q-1})$, subject to the constraints $0 < x_0, ..., x_{q-1} \le t$ and $\sum_{i=0}^{q-1} x_i = k$, is obtained at $x_0 = \cdots = x_{q-1} = \frac{k}{q}$.
proof: By Prop. $10_{(c)}$, $g(x_0, x_1, ..., x_{q-1}) = \sum_{i=0}^{q-1} x_i^2 - (t+1) \cdot k + \frac{1}{2}t(t+1) \cdot q$ (Use $0 < x_i \le t$). The function $g(\cdot, \cdot, \cdots, \cdot)$ is a quadratic form in the x_i's. □

Notation:

$$h_n^\rho(q) \stackrel{\text{def}}{=} q \cdot (f(\frac{\rho n}{q}) + f(\frac{n - \rho n}{q}))$$

Proposition 12: Let Q be the set of integers q, satisfying $\frac{\rho n}{t} \le q \le n - \rho n$. Then

$$C_{MR}(n, \rho, t) \ge \frac{1}{nt} \cdot \min_{q \in Q}\{h_n^\rho(q)\}$$

proof: Immediate by combining Prop. $10_{(a)}$ and 11, using the fact that $r \in MR_n^\rho$ contains no long homogeneous substrings. □

Proposition 13:

$$h_n^\rho(q) = \frac{t+1}{n} \cdot q + \frac{(1 + 2\rho^2 - 2\rho)n}{t} \cdot \frac{1}{q} - \frac{t+1}{t}.$$

The minimum of the function $h_n^\rho(\cdot)$, over $q \in Q$, is obtained at:

$$q_{min} \stackrel{\text{def}}{=} \sqrt{\frac{1 + 2\rho^2 - 2\rho}{t(t+1)}} \cdot n$$

The minimum value, $h_n^\rho(q_{min})$, is:

$$v_t^\rho \stackrel{\text{def}}{=} \sqrt{(4 + 8\rho^2 - 8\rho) \cdot \frac{t+1}{t}} - \frac{t+1}{t}$$

Combining Prop. 12 and 13, we get

Lemma 3: $\quad C_{MR}(n, \rho, t) \ge v_t^\rho$.

3.3 The Lower Bound Theorem

Combining Lemmas 1, 2 and 3, we get

Theorem 1: $C(n, \rho, t)$ is at least

$$(\sqrt{(2 + 8(\rho - \frac{1}{2})^2) \cdot \frac{t+1}{t}} - \frac{t+1}{t}) - \frac{3t}{n}$$

4. UPPER BOUND ON $C(n,\rho,t)$

In this section we demonstrate the tightness of the lower bound presented above. Namely,

Theorem 2: $C(n,\rho,t)$ is at most

$$\left(\sqrt{(2+8(\rho-\frac{1}{2})^2)\cdot\frac{t+1}{t}}-\frac{t+1}{t}\right)+\frac{t+1}{n}+\frac{1}{2t^2}$$

proof: The Theorem follows from observing that the proof of the lower bound specifies the structure of a string which achieves minimum $c_t(\cdot)$ among all strings in MR_n^ρ. The only problem in constructing such a string is that non-integer numbers, of blocks and block sizes, may appear. The reader can easily verify that the "overline" added by rounding-up the number of blocks and their sizes is less than $\frac{t+1}{n}$ and $\frac{1}{2t^2}$, respectively. For more details, see our technical report [1]. □

5. CONCLUSIONS

The reader may easily verify that the gap between the lower and upper bounds is $O(\frac{1}{t}+\frac{t}{n})$. Let us approximate the expressions given by the Theorems, ignoring these additive error terms. We get

[a] $C(n,\frac{1}{2},t)\approx\sqrt{2}-1\approx 0.414$

[b] $C(n,0.676,t)<\frac{1}{2}$.

[c] $C(n,0.677,t)>\frac{1}{2}$.

ACKNOWLEDGMENTS

I am indebted to Tom Leighton for teaching me how to count (overlaps). It is my pleasure to thank Michael Ben-Or, Benny Chor, Shafi Goldwasser, Hans Heller, Silvio Micali, Gary Miller, Ron Rivest and Avi Wigderson for very helpful discussions, useful ideas and consistent encouragement. I would also like to thank the referees for their suggestions.

REFERENCES

[1] Goldreich, O., "On the Number of Close-and-Equal Pairs of Bits in a String (with Implications on the Security of RSA's L.S.B)", MIT/LCS/TM-256, March 1984.