Free Energy Minimization Algorithm for Decoding and Cryptanalysis

David J.C. MacKay
Cavendish Laboratory, Cambridge CB3 0HE
mackay@mrao.cam.ac.uk

Submitted to *Electronics Letters* November 23, 1994; published 16th March 1995 (vol. 31 no.6)

Abstract

An algorithm is derived for inferring a binary vector $\mathbf{s}$ given noisy observations of $\mathbf{As}$ modulo 2, where $\mathbf{A}$ is a binary matrix. The binary vector is replaced by a vector of probabilities, optimized by free energy minimization. Experiments on the inference of the state of a linear feedback shift register indicate that this algorithm supersedes Meier and Staffelbach's polynomial algorithm.

Index: approximate inference, combinatorial optimization, stream cipher.

Consider three binary vectors: $\mathbf{s}$ of length N, and $\mathbf{z}$ and $\mathbf{n}$ of length $M \geq N$, related by:

$$(\mathbf{As} + \mathbf{n}) \bmod 2 = \mathbf{z} \tag{1}$$

where $\mathbf{A}$ is a binary matrix. Our task is to infer $\mathbf{s}$ given $\mathbf{z}$ and $\mathbf{A}$, and given assumptions about the statistical properties of $\mathbf{s}$ and $\mathbf{n}$. This problem arises in the decoding of a noisy signal transmitted using a linear code $\mathbf{A}$, and in the inference of the sequence of a linear feedback shift register (LFSR) from noisy observations [1, 2].

I assume that the prior probability distribution of $\mathbf{s}$ and $\mathbf{n}$ is separable thus: $P(\mathbf{s}, \mathbf{n}) = \prod_n P(s_n) \prod_m P(n_m)$. The log probability of $\mathbf{z}$ as a function of $\mathbf{s}$ can be written in terms of the noise free vector $\mathbf{t}(\mathbf{s}) = \mathbf{As} \bmod 2$:

$$\log P(\mathbf{z}|\mathbf{s}, \mathbf{A}) \quad = \quad \sum_m t_m(\mathbf{s})\, g_m + \text{const.} \tag{2}$$

where $g_m \equiv \log[P(n_m=1)/P(n_m=0)]$ if $z_m=0$ and $g_m \equiv -\log[P(n_m=1)/P(n_m=0)]$ if $z_m=1$. The posterior distribution of $\mathbf{s}$ is, by Bayes' theorem:

$$P(\mathbf{s}|\mathbf{z},\mathbf{A}) = \frac{P(\mathbf{z}|\mathbf{s},\mathbf{A})P(\mathbf{s})}{P(\mathbf{z}|\mathbf{A})}. \tag{3}$$

I assume our aim is to find the most probable $\mathbf{s}$, but that an exhaustive search over all 2^N possible sequences $\mathbf{s}$ is not feasible. One way to attack such a combinatorial optimization problem is via a related continuous problem in which the discrete variables are replaced by real variables [3]. Here I derive a continuous representation in terms of a free energy approximation [4]. I approximate the awkward probability distribution (3) by a simpler separable distribution $Q(\mathbf{s};\theta) \equiv \prod_n q_n(s_n;\theta_n)$, parameterized thus:

$$q_n(s_n=1;\theta_n) = \frac{1}{1+e^{-\theta_n}} \equiv q_n^1; \;\; q_n(s_n=0;\theta_n) = 1-q_n^1 \equiv q_n^0. \tag{4}$$

The parameters θ are adjusted to find a θ^* that minimizes the variational free energy,

$$F(\theta) = \sum_{\mathbf{s}} Q(\mathbf{s};\theta) \log \frac{Q(\mathbf{s};\theta)}{P(\mathbf{z}|\mathbf{s},\mathbf{A})P(\mathbf{s})}, \tag{5}$$

the hope being that the $\mathbf{s}$ that maximizes $Q(\mathbf{s};\theta^*)$ may also maximize $P(\mathbf{s}|\mathbf{z},\mathbf{A})$. Although F is defined by a summation over the 2^N discrete values of $\mathbf{s}$, it is possible to evaluate F and its gradient $\partial F/\partial\theta$ in a time that is proportional to the weight of $\mathbf{A}$, $w_{\mathbf{A}}$ (*i.e.*, the number of ones in $\mathbf{A}$), as will now be shown.

F separates into three terms, $F(\theta) = E_L(\theta) + E_P(\theta) - S(\theta)$, where the 'entropy' is: $S(\theta) \equiv -\sum_{\mathbf{s}} Q(\mathbf{s};\theta)\log Q(\mathbf{s};\theta) = -\sum_n \left[q_n^0 \log q_n^0 + q_n^1 \log q_n^1\right]$, with derivative: $\frac{\partial}{\partial\theta_n} S(\theta) = -q_n^0 q_n^1 \theta_n$; the 'prior energy' is: $E_P(\theta) \equiv -\sum_{\mathbf{s}} Q(\mathbf{s};\theta)\log P(\mathbf{s}) = -\sum_n b_n q_n^1$ where $b_n = \log[P(s_n=1)/P(s_n=0)]$, and has derivative $\frac{\partial}{\partial\theta_n} E_P(\theta) = -q_n^0 q_n^1 b_n$; and the 'likelihood energy' is:

$$E_L(\theta) \equiv -\sum_{\mathbf{s}} Q(\mathbf{s};\theta)\log P(\mathbf{z}|\mathbf{s},\mathbf{A}) = -\sum_m g_m \sum_{\mathbf{s}} Q(\mathbf{s};\theta)\, t_m(\mathbf{s}) + \text{const.} \tag{6}$$

We can compute $\sum_{\mathbf{s}} Q(\mathbf{s};\theta)\, t_m(\mathbf{s})$ for each m by a 'forward' recursion involving a sequence of probabilities $p_{m,\nu}^1$ and $p_{m,\nu}^0$ for $\nu = 1 \ldots N$, defined to be the probabilities that the partial sum $t_m^{1\nu} = \sum_{n=1}^{\nu} A_{mn} s_n \bmod 2$ is equal to 1 and 0 respectively. These probabilities satisfy:

$$\left.\begin{array}{rcl} p_{m,\nu}^1 &=& q_\nu^0 p_{m,\nu-1}^1 + q_\nu^1 p_{m,\nu-1}^0 \\ p_{m,\nu}^0 &=& q_\nu^0 p_{m,\nu-1}^0 + q_\nu^1 p_{m,\nu-1}^1 \end{array}\right\} \text{ if } A_{m\nu}=1\text{; else } \left\{\begin{array}{rcl} p_{m,\nu}^1 &=& p_{m,\nu-1}^1 \\ p_{m,\nu}^0 &=& p_{m,\nu-1}^0 \end{array}\right. , \tag{7}$$

with initial condition $p_{m,0}^1 = 0, p_{m,0}^0 = 1$. We obtain: $E_L(\theta) = -\sum_m g_m p_{m,N}^1$. The derivative of E_L with respect to θ_n can be obtained by evaluating for each m a 'reverse' sequence of probabilities $r_{m,\nu}^1$ and $r_{m,\nu}^0$, defined to be the probabilities that the partial

sum $t_m^{\nu N} = \sum_{n=\nu}^{N} A_{mn} s_n \bmod 2$ is equal to 1 and 0 respectively. Then using the relation $p^1_{m,N} = q^0_n \left(p^1_{m,n-1} r^0_{m,n+1} + p^0_{m,n-1} r^1_{m,n+1} \right) + q^1_n \left(p^1_{m,n-1} r^1_{m,n+1} + p^0_{m,n-1} r^0_{m,n+1} \right)$ and defining $d_{mn} = \left(p^1_{m,n-1} r^1_{m,n+1} + p^0_{m,n-1} r^0_{m,n+1} \right) - \left(p^1_{m,n-1} r^0_{m,n+1} + p^0_{m,n-1} r^1_{m,n+1} \right)$, we obtain the derivative $\frac{\partial}{\partial \theta_n} E_L(\theta) = -q^0_n q^1_n \sum_m g_m d_{mn}$. Thus the derivative of the free energy is:

$$\frac{\partial F}{\partial \theta_n} = q^0_n q^1_n \left[\theta_n - b_n - \sum_m g_m d_{mn} \right]. \tag{8}$$

The assignment:

$$\theta_n := b_n + \sum_m g_m d_{mn} \tag{9}$$

sets the derivative to zero and is guaranteed to reduce the free energy. This re-estimation equation can be efficiently interleaved with the reverse recursion, giving a simple optimizer of F. Optimizers of F can be modified by using 'deterministic annealing' [5], in which the non-convexity of the objective function F is switched on gradually by varying an 'inverse temperature' β from 0 to 1. This procedure is intended to prevent the algorithm from running into the local minimum that the initial gradient points towards. We define $F(\theta, \beta) = \beta E_L(\theta) + E_P(\theta) - S(\theta)$, and perform a sequence of minimizations of this function over θ with successively larger values of β.

The success of the algorithm is expected to depend on the representation of $\mathbf{s}$, with best results if $\mathbf{A}$ is sparse and the true posterior distribution over $\mathbf{s}$ is close to separable.

Computational complexity: The algorithm is expected to take of order 1, or at most N, gradient evaluations to converge, so that the total time taken is of order between $w_{\mathbf{A}}$ and $w_{\mathbf{A}} N$. Memory proportional to $w_{\mathbf{A}}$ is required.

Cryptanalysis application

Various demonstrations of this algorithm are given in [6]. Here I describe an application to a cryptanalysis problem, building on the method of Meier and Staffelbach [1]. Assume a LFSR of length k bits with t taps produces a sequence $\mathbf{a}_0$ of length N bits, and noisy observations $\mathbf{a}_1 = (\mathbf{a}_0 + \mathbf{s}) \bmod 2$ are made (for details see [1],[2]). Here $\mathbf{s}$ is a sparse noise vector of length N. For $N \gg k$, as in ref. [1], we can create a sparse $M \times N$ matrix $\mathbf{A}$ of parity checks such that $\mathbf{A}\mathbf{a}_0 \bmod 2 = 0$, each row of $\mathbf{A}$ having weight $(t+1)$. The noisy sequence $\mathbf{a}_1$ violates some of these parity checks as described by the vector $\mathbf{z} \equiv \mathbf{A}\mathbf{a}_1$. Then our problem is to find the noise vector $\mathbf{s}$ that satisfies:

$$\mathbf{A}\mathbf{s} \bmod 2 = \mathbf{z}, \tag{10}$$

and that has maximum prior probability, given our knowledge of the noisy observation process. [There are many (2^k) values of $\mathbf{s}$ satisfying equation (10), one for each of the possible initial LFSR states.] In (10), unlike (1), there is no noise added to $\mathbf{As}$. However, we can apply the free energy method to a sequence of problems of the form $(\mathbf{As} + \mathbf{n}) \bmod 2 = \mathbf{z}$ with increasing inverse temperature β, such that the noise–free task is the limiting case, $\beta = \infty$.

Experimental results

Test data were created for specified k and N using random taps in the LFSR and random observation noise with fixed uniform probability. The parameter β was initially set to 0.25. For each value of β, the optimization was run until the decrease in free energy was below a specified tolerance (0.001). β was increased by factors of 1.4 until either the most probable vector under $Q(\mathbf{s}; \theta)$ satisfied (10), or until a maximum value of $\beta = 4$ was passed.

Figure 1 here.

Results are shown in figure 1. Each dot represents an experiment. A box represents a successful decoding. On each graph a horizontal line shows an information theoretic noise bound above which one does not expect to be able to infer $\mathbf{s}$, and two curved lines, from tables 3 and 5 of ref. [1], show (lower line) the limit up to which Meier and Staffelbach's 'algorithm B appeared to be very successful in most experiments' and (upper line) the theoretical bound beyond which their approach is definitely not feasible.

Conclusion

This paper has derived an algorithm with a well-defined objective function for inference problems in modulo 2 arithmetic. In application to a cryptanalysis problem, this algorithm is similar to Meier and Staffelbach's [1] algorithm B and thus answers their question of whether a derivation could be provided. But it is not identical: the details of the mapping from $[0,1]^N \rightarrow [0,1]^N$ are different, and there is no analogue of their multiple 'rounds' in which the data vector $\mathbf{a}_1$ is changed. The new algorithm appears to give superior performance and frequently succeeds at parameter values right up to the upper theoretical limits derived by Meier and Staffelbach.

Acknowledgements

I thank R. Anderson, R. Neal and R. Sewell for helpful discussions.

References

[1] W. Meier and O. Staffelbach. Fast correlation attacks on certain stream ciphers. *J. Cryptology*, 1:159–176, 1989.

[2] R.J. Anderson. Searching for the optimum correlation attack. In B. Preneel, ed., *Proceedings of 1994 K.U. Leuven Workshop on Cryptographic Algorithms*, Lecture Notes in Computer Science. Springer-Verlag, 1995.

[3] J.J. Hopfield and D.W. Tank. Neural computation of decisions in optimization problems. *Biological Cybernetics*, 52:1–25, 1985.

[4] R.P. Feynman. *Statistical Mechanics*. W. A. Benjamin, Inc., 1972.

[5] D.E. Van den Bout and T.K. Miller III. Improving the performance of the Hopfield-Tank neural network through normalization and annealing. *Biological Cybernetics*, 62:129–139, 1989.

[6] D.J.C. MacKay. A free energy minimization framework for inference problems in modulo 2 arithmetic. In *Proceedings of 1994 K.U. Leuven Workshop on Cryptographic Algorithms*, 1995.

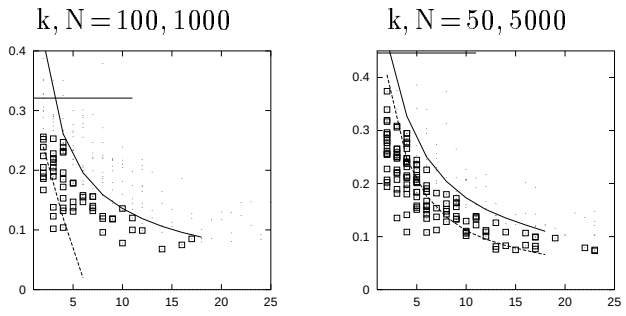

Figure 1: Results for cryptanalysis problem as a function of number of taps (horizontal axis) and noise level (vertical).