Dispersion and Diffusion

Samer Seraj
University of Toronto

Abstract. In Shannon’s landmark 1949 paper *Communication Theory of Secrecy Systems*, the idea of diffusive functions is briefly mentioned. We consider two common definitions of such functions mapping between binary strings of fixed length. Given the dimension of the input space, we determine the minimum dimension of the output space for which such a function exists, by explicit construction and with respect to each definition. It will follow that each larger output dimension allows for such a function as well.

It has been noted in cryptography conferences and workshops that diffusion is a desirable property for certain functions. However, what is diffusion precisely? Shannon [1] says

“In the method of diffusion the statistical structure of $M$ which leads to its redundancy is “dissipated” into long range statistics - i.e. into statistical structure involving long combinations of letters in the cryptogram. The effect here is that the enemy must intercept a tremendous amount of material to tie down this structure, since the structure is evident only in blocks of very small individual probability. Furthermore, even when he has sufficient material, the analytical work required is much greater since the redundancy has been diffused over a large number of individual statistics.”

Our functions will map between fixed length strings of 0s and 1s. One notion is that if a bit is flipped (0 to 1, or 1 to 0) in any fixed input, then some half of the bits in the corresponding output flips; we call this dispersion. A second concept is that if a bit is flipped in any fixed input, then, for each bit in the corresponding output, the probability of it flipping is half; we accept this as diffusion. Some make no distinction between the two, but we will see that the terms are not interchangeable.

Definition. For each $n \in \mathbb{Z}^+$:

1. The set of $n$-bit binary strings are $n$-tuples of elements from $\mathbb{F}_2 = \{0, 1\}$, denoted by $\mathbb{F}_2^n$. The bits of each $x \in \mathbb{F}_2^n$ are indexed from 1 to $n$, from left to right.

2. The XOR binary operation $\oplus$ on $\mathbb{F}_2^n$ is defined as bitwise “addition” in the field $\mathbb{F}_2$. A generalization to arbitrary pairs of finite bit strings is that 0s are first appended to the left of the shorter string to force the same length.

3. The Hamming weight $w : \mathbb{F}_2^n \to \mathbb{Z}$ of $x \in \mathbb{F}_2^n$ is the number of non-zero bits of $x$.

4. The Hamming distance $h : \mathbb{F}_2^n \times \mathbb{F}_2^n \to \mathbb{Z}$ between $x, y \in \mathbb{F}_2^n$ is the number of bits on which they disagree, so $h(x, y) = w(x \oplus y)$.

5. An injective function $f : \mathbb{F}_2^n \to \mathbb{F}_2^m$ is said to be dispersive if $m$ is even and

$$\forall x, y \in \mathbb{F}_2^n : h(x, y) = 1 \implies h(f(x), f(y)) = \frac{m}{2}.$$  

6. For each $i \in \mathbb{Z}$ such that $1 \leq i \leq n$, let $\pi_i : \mathbb{F}_2^n \to \mathbb{Z}$ be defined by taking $i$th bit of the input, and mapping 0,1 $\in \mathbb{F}_2$ to 0,1 $\in \mathbb{Z}$ respectively.

7. Let $E_n = \{ \{x, y\} \subseteq \mathbb{F}_2^n : h(x, y) = 1 \}$, which is all pairs at a distance of 1, so it is our sample space. It follows from elementary combinatorial reasoning that the cardinality of $E_n$ is $n2^{n-1}$.

1samer.seraj@mail.utoronto.ca
8. An injective function \( g : \mathbb{F}_2^n \to \mathbb{F}_2^m \) is said to be diffusive if, for each integer \( 1 \leq i \leq n \), we have
\[
\frac{1}{n2^{n-1}} \cdot \sum_{(x,y) \in E_n} \pi_i(g(x) \oplus g(y)) = \frac{1}{2}.
\]
We first optimize dispersive maps, with regards to minimizing the output dimension.

**Definition.**

7. A subset of \( \mathbb{F}_2^n \) is said to be a code. An element of a code is a codeword.

8. A code is said to be linear if it is closed under the \( \oplus \) binary operation. It is easy to see that a vector space is formed by a linear code over the field \( \mathbb{F}_2 \).

9. Let \( n \in \mathbb{Z}^+ \) be even. Then \( x \in \mathbb{F}_2^n \) is said to be semi-weight if \( w(x) = \frac{n}{2} \).

**Proposition 1.** For positive integers \( n \equiv 0, 1, 2, 3 \pmod{4} \) respectively, the minimum \( m \in \mathbb{Z}^+ \) for which there exists a dispersive map \( f_n : \mathbb{F}_2^n \to \mathbb{F}_2^m \) is \( m = n + 2, n + 1, n + 1 \). Every output dimension greater than the minimum allows for dispersive functions as well.

*Proof.* For each \( n \in \mathbb{Z}^+ \), \( \mathbb{F}_2^n \) has \( 2^n \) elements, so in order for an injective function \( f_n : \mathbb{F}_2^n \to \mathbb{F}_2^m \) to exist, \( m \geq n \). We will show that \( m = n \) can be achieved for \( n \equiv 2 \pmod{4} \). For odd \( n \), since we always require even \( m \), we need \( m \geq n + 1 \) which will also be proven to be achievable.

However we will show that, for \( n \equiv 0 \pmod{4} \), no permutation of \( \mathbb{F}_2^n \) is dispersive, so we will instead find a map to \( \mathbb{F}_2^{n+2} \). For suppose \( m = n \). Let \( x = x_1 \cdots x_n \in \mathbb{F}_2^n \). Suppose its non-zero bits are at indices \( i_1, \ldots, i_k \). Consider the \( k \)-tuple of vectors \( (v_1, \ldots, v_k) \) where in \( v_i \) the non-zero bits are precisely those at indices \( i_1, \ldots, i_j \). Assume without loss of generality that 0 maps to 0, since \( (x \oplus z) \oplus (y \oplus z) = x \oplus y \). Note that XOR-ing two even weight vectors produces another vector of even weight. Since \( f_n(0) = 0 \) is even weight, and each of \( f_n(v_i) \oplus f_n(v_{i+1}) \) is semi-weight and so even weight, each \( f_n(v_i) \) is even weight. Thus \( v_k = x \) (and so every vector) must map to an even-weight vector, of which there are only \( 2^{n-1} < 2^n \).

The Hamming cube \( \mathbb{F}_2^n \) is a linear code which can be generated by the row vectors \( (e_1, \ldots, e_n) \) of the \( n \times n \) identity matrix. For \( k \equiv 0, 2 \pmod{4} \) respectively, we will find \( k - 1 \) and \( k \) linearly independent semi-weight vectors \( v_i \in \mathbb{F}_2^k \) and apply them in the following way. Recall that each element of \( \mathbb{F}_2^n \) can be uniquely represented as a sum of the \( e_i \). Define
\[
f_n \left( \sum_{i=1}^{n} a_i \cdot e_i \right) = \sum_{i=1}^{n} a_i \cdot v_{i,m},
\]
where \( m = n + 2, n + 1, n + 1 \) respectively for \( n \equiv 0, 1, 2, 3 \pmod{4} \), and the \( a_i \) are from \( \mathbb{F}_2 \). The map \( f_n \) is injective due to the linear independence of the \( v_{i,m} \), and the semi-weight property of the \( v_{i,m} \) means that a 1-bit change in the input causes a change in some half of the bits in the output.

Finally, we define the \( v_i \) (the subscript under the subscript is dropped now for cleaner text). The identity matrix has linearly independent rows \( (e_1, \ldots, e_k) \), so the elementary operation of adding one row to another with our vector addition \( \oplus \) preserves linear independence. The case of \( k = 2 \) is trivial as the identity map works. For even integers \( k \geq 4 \), we define
\[
v_i = \begin{cases} \sum_{j=1}^{i-1} e_j & 1 \leq i \leq \frac{k}{2} \\ e_i \oplus v_{\frac{k}{2}} \oplus e_k & \frac{k}{2} + 1 \leq i \leq k - 1 \\ e_k \oplus v_{1} \oplus v_{\frac{k}{2}+1} & i = k. \end{cases}
\]
The final one is only for $k \equiv 2 \pmod{4}$. Since the $v_i$ are all semi-weight, the construction is complete. It is evident that the dimension of the output space can be chosen to be any even integer greater than $m$, if desired, by using some of the $v_i$ from the higher dimension.

In the $n \equiv 2 \pmod{4}$ case, if the column vectors are used as the $v_i$ instead of the rows, then for each bit in the output, a change in one of exactly some half of the input bits causes that output bit to flip. Note that, since column rank equals row rank, linear independence of the rows implies that of the columns. This produces a diffusive map, but unfortunately the matrices are not square in the other three cases. Now we optimize diffusive functions, again with respect to minimizing the output dimension. Incredibly, we will see that there exists a diffusive permutation of every input space $F_2^n$ for integers $n \geq 2$.

**Definition.**

10. For bit strings $x \in F_2^n$ and $y \in F_2^m$, denote by $x|y$ the concatenation in $F_2^{n+m}$ of $x$ to the left of $y$.

11. For each bit string $x \in F_2^n$ and integer $n \geq 2$, let $\alpha(x)$ denote the leftmost bit of $x$ as an element of $F_2$. Similarly, let $\beta(x)$ denote the string in $F_2^{n-1}$ when leftmost bit of $x$ is removed.

12. Let $\tau : F_2^n \to F_2^n$ denote the function which transposes the rightmost two bits. Consequently define $\sigma : F_2^n \to F_2^n$ by $\sigma(x) = \tau(x \oplus 1)$, where we use the generalized definition of $\oplus$. It will be useful to note that $\beta \circ \sigma = \sigma \circ \beta$ for integers $n \geq 3$.

13. The complement $\bar{x}$ of a codeword $x \in F_2^n$ replaces each 0 in $x$ with 1, and each 1 in $x$ with 0.

**Proposition 2.** For each integer $n \geq 2$, there exists a diffusive function $g_n : F_2^n \to F_2^n$ and for every larger output dimension $m \geq n$. Of course we cannot have the output dimension less than $n$ due to injectivity.

**Proof.** It is straightforward to verify that for $n = 2$, the identity function $g_2 : F_2^2 \to F_2^2$, defined by $g_2(x) = x$, is diffusive. We recursively define $g_n$ for integers $n \geq 3$ as

$$ g_n(x) = \begin{cases} (\alpha \circ g_{n-1} \circ \beta)(x)(g_{n-1} \circ \beta)(x) & \alpha(x) = 0 \\ (\alpha \circ g_{n-1} \circ \beta \circ \sigma)(x)(g_{n-1} \circ \beta \circ \sigma)(x) & \alpha(x) = 1. \end{cases} $$

We must prove that $g_n$ is bijective and satisfies the diffusion property that for every integer $1 \leq i \leq n$,

$$ \sum_{\{x,y\} \in E_n} \pi_i(g_n(x) \oplus g_n(y)) = n2^{n-2}. $$

Both will be proven by induction with $g_2$ as the basis, which we have previously mentioned is easily seen to be bijective and satisfy diffusion. Then suppose that for some integer $n-1 \geq 2$, it holds that $g_{n-1}$ is bijective and satisfies the diffusion equation above.

$\sigma$ is easily seen to be a bijection of $F_2^n$ because, for each $x \in F_2^{n-2}$, it cycles

$$ x|00 \to x|10 \to x|11 \to x|01 \to x|00; $$

this is also important in the latter half of the proof. Then

$$ g_n(F_2^n) = \{ (\alpha \circ g_{n-1})(x) \mid g_{n-1}(x) : x \in F_2^{n-1} \} \cup \{ (\alpha \circ g_{n-1} \circ \sigma)(x) \mid g_{n-1}(x) : x \in F_2^{n-1} \} $$

$$ = \{ (\alpha \circ g_{n-1})(x) \mid g_{n-1}(x) : x \in F_2^{n-1} \} \cup \{ (\alpha \circ g_{n-1})(x) \circ g_{n-1}(x) : x \in F_2^{n-1} \} $$

$$ = \{ 0,0 \mid x : x \in F_2^{n-1} \} = F_2^n. $$

For the diffusive property, first note that for each integer $1 \leq i \leq n$,

$$ \sum_{\{x,y\} \in E_n} \pi_i(g_n(x) \oplus g_n(y)) = \sum_{\phi=0,1} \left( \sum_{\{x,y\} \in E_{n-1}} \pi_i(g_n(\phi x) \oplus g_n(\phi y)) \right) + \sum_{x \in F_2^{n-1}} \pi_i(g_n(0|x) \oplus g_n(1|x)). $$
From left to right, name these four sums $c_i, p_i, q_i, r_i$ ($p_i$ for $\phi = 0$ and $q_i$ for $\phi = 1$). We have $p_1 = p_2$ since the leftmost bit in each term is simply a copy of the second-leftmost bit. Similarly $q_1 = q_2$ as the leftmost bit in each term is the complement of the second-leftmost bit and $\bar{x} \oplus \bar{y} = x \oplus y$. So in either case, we may assume without loss of generality that $2 \leq i \leq n$. Then

$$p_i = \sum_{\{x, y\} \in E_{n-1}} \pi_i(g_n(0|x) \oplus g_n(0|y)) = \sum_{\{x, y\} \in E_{n-1}} \pi_{i-1}(g_{n-1}(x) \oplus g_{n-1}(y)) = (n-1)2^{n-3},$$

by the induction hypothesis. Since $\sigma$ is bijective on $\{0|x : x \in \mathbb{F}_2^{n-1}\}$ and we have the identities

$$\begin{cases} 
  g_n(1|x) \oplus g_n(1|y) = (g_n \circ \sigma)(0|x) \oplus (g_n \circ \sigma)(0|y) \\
  h(\sigma(x), \sigma(y)) = h(x, y)
\end{cases}$$

immediately from definitions, it follows that

$$q_i = \sum_{\{x, y\} \in E_{n-1}} \pi_i(g_n(1|x) \oplus g_n(1|y)) = \sum_{\{x, y\} \in E_{n-1}} \pi_i((g_n \circ \sigma)(0|x) \oplus (g_n \circ \sigma)(0|y))$$

$$= \sum_{\{x, y\} \in E_{n-1}} \pi_i(g_n(0|x) \oplus g_n(0|y)) = p_i = (n-1)2^{n-3}.$$

Using the fact that $x \oplus \bar{y} = \bar{x} \oplus y$,

$$r_i = \sum_{x \in \mathbb{F}_2^{n-1}} \pi_1(g_n(0|x) \oplus g_n(1|x)) = \sum_{x \in \mathbb{F}_2^{n-1}} \pi_1((\alpha \circ g_{n-1})(x) \oplus (\alpha \circ g_{n-1} \circ \sigma)(x))$$

$$= \sum_{x \in \mathbb{F}_2^{n-1}} \pi_1(g_{n-1}(x) \oplus (g_{n-1} \oplus \sigma)(x)) = 2^{n-1} - r_2,$$

since there are $2^{n-1}$ terms in the sum and the above is clearly the complement of $r_2$. We will prove that $r_2 = 2^{n-2}$, which will result in $r_1 = 2^{n-1} - r_2 = 2^{n-2}$, so assume without loss of generality that $2 \leq i \leq n$. Then

$$r_i = \sum_{x \in \mathbb{F}_2^{n-1}} \pi_i(g_n(0|x) \oplus g_n(1|x)) = \sum_{x \in \mathbb{F}_2^{n-1}} \pi_{i-1}(g_{n-1}(x) \oplus (g_{n-1} \oplus \sigma)(x)).$$

Recalling how the permutation representation of $\sigma$ is the product of all 4-cycles $(x00 \ x10 \ x11 \ x01)$, the above sum can be written as the sum over all $x \in \mathbb{F}_2^{n-3}$ of the quadruple sum

$$\pi_{i-1}(g_{n-1}(x|00) \oplus g_{n-1}(x|10)) + \pi_{i-1}(g_{n-1}(x|10) \oplus g_{n-1}(x|11)) + \pi_{i-1}(g_{n-1}(x|11) \oplus g_{n-1}(x|01)) + \pi_{i-1}(g_{n-1}(x|01) \oplus g_{n-1}(x|00)).$$

It is a standard induction argument to prove that each such quadruple sum is equal to 2 for $2 \leq i \leq n$, so the sum over all such quadruples is $2 \cdot 2^{n-3} = 2^{n-2}$. Therefore

$$c_i = p_i + q_i + r_i = (n-1)2^{n-3} + (n-1)2^{n-3} + 2^{n-2} = n2^{n-2},$$

as desired. Certainly by appending the leftmost bit of each output to that output, a diffusive injection into an output space of larger dimension can be constructed.

There are a number of natural questions that arise from our discourse. For example, one may consider maps which are both dispersive and diffusive.

Another idea is to call an injective function $f : \mathbb{F}_2^n \to \mathbb{F}_2^n$ $k$-dispersive if $m$ is even and

$$\forall x, y \in \{0, 1\}^n : 1 \leq h(x, y) \leq k \implies h(f(x), f(y)) = m/2.$$

4
We have explored $k = 1$, and while other small $k \geq 2$ seem difficult to optimize, bounds on the minimal dimensions and the construction of near-optimal maps would be an accomplishment.

One may similarly define and explore $k$-diffusive maps, where the sample space would be

$$E_{n,k} = \{ \{x, y\} \subseteq \mathbb{F}_2^n : 1 \leq h(x, y) \leq k \},$$

instead of the $E_n$. In this framework, we have solved $k = 1$, and work on larger $k$ would be of interest.

References

1. Shannon, (1949) *Communication Theory of Secrecy Systems*, Bell System Technical Journal, Vol. 28, pp. 656-715