Estimating Distortion Consistently : An Algebraic Approach

by Gordon Simons[1]

University of North Carolina at Chapel Hill

Abstract

It is argued that it is feasible to estimate the extent of distortion of a distorted image *before* attempting to reconstruct the image. One might choose to use such an estimate to help one decide whether a reconstruction should be attempted. Or the estimate might reasonably influence the method of reconstruction chosen. A general model is described under which a consistent estimator is obtainable. A proof of consistency is given, and the methodology is used on an intentionally distorted natural language text. The methods, which are primarily algebraic, are loosely motivated by an established method of encryption.

Short title: Estimating distortion
American Mathematical Society 1980 subject classifications. Primary: 62F12; Secondary: 62F10, 62G05, 60B15
Key words and phrases: Distortion, consistent estimators, entropy, binary bits, convolutions on groups

[1]The author's research was partially supported by the National Science Foundation, Grant No. DMS–8701201.

1. INTRODUCTION AND SUMMARY. Consider a complex structure such as a written text, a visual image or human speech, that has been subjected to a certain amount of distortion in the process of communication. The problem under consideration is that of quantifying and estimating the level of distortion. While the issue of reconstruction has properly commanded considerable attention in recent years, the problem of quantitatively assessing the distortion appears to have been neglected. Such an assessment could have a bearing on how, or whether, a reconstruction should be attempted. This presumes, as we do here, that it is feasible to estimate the distortion without simultaneously attempting a reconstruction.

While the problem of estimating distortion might reasonably be approached from the perspective of entropy, we will, instead, exploit an interesting algebraic structure associated with a secure method of encryption — one commonly used for communicating with embassies. Because of this connection, and mainly because of an easy access to computer files with various kinds of written text, this paper will emphasize distortion of natural language texts. However, it seem likely that the methodology developed here will be most useful in other settings, wherever distortion is a serious problem.

Let $X_1, X_2, \cdots$ denote a natural language text, called "plaintext", which, for convenience, has been converted to a binary format — zeros and ones. It is customary to refer to the original uncoded text (commonly in the Roman alphabet) as the "plaintext". We shall call the X sequence, when clarity requires, the "plaintext sequence". Further, let $Y_1, Y_2, \cdots$be a sequence of independent Bernoulli random variables with a common mean p, and form a new sequence $Z_1, Z_2, \cdots$ by addition modulo two:

$$Z_i = (X_i + Y_i) \bmod 2, \; i = 1,2, \cdots. \tag{1}$$

A completely secure encryption is obtained by using $p = \frac{1}{2}$. This causes the Z sequence, which can be transmitted through an unprotected channel, to be iid Bernoulli random variables with mean one–half, i.e., pure noise. Careful security is required for the Y sequence, the "key". For, with it, the original message is easily recovered:

$$X_i = (Z_i + Y_i) \bmod 2, i = 1,2, \cdots. \tag{2}$$

The Y sequence will be used to model distortion. The Bernoulli parameter p is assumed to be unknown. The effects of various distortion levels on a well–known quotation are illustrated in Table 1.

p	Distortion of Quotation	p	Distortion of Quotation
.00	BEWARE THE IDES OF MARCH.	.10	BEWBRE THE BDET OHPJARCG.
.01	BEWARE THE IDES OF MARCH.	.20	NIWAQE TDE ILPSMMZXMYVDDU
.02	BGXARE,THE IDES OF MARCE.	.30	YM IBAXTU.XG?CYPIQMPATGS.
.03	BEWARE'THE IDES OF MGTCHY	.40	NQ?PJN?RECJJU.ET""IBDRBZY
.05	BAWARG THE IDES PE,MARCH.	.50	,LKB.QBLXOBF WVBMCZKXBDIX

Table 1: Distortions of a Widely Known Quotations

Each of the 26 letters of the alphabet, together with five punctuation marks and the delimiter between words (a "space"), is assigned a different five–bit code. Each bit is reversed with probability p.

The extent of the distortion is best viewed in terms of the distance of p from one–half. Clearly, p = 0 represents no distortion; p = 1 represents a complete *but systematic* distortion which is equivalent to recoding the X sequence, with ones and zeros reversed, and replacing p = 1 by p = 0. By such reasoning, we are lead to view the values of p and 1–p as yielding *equivalent* amounts of distortion. A simple definition of *distortion*, which reflects this viewpoint, the one we shall use, is provided by the notion of entropy (see, for instance, Billingsley (1965), page 60):

$$\theta = -p \cdot \log_2 p - (1-p) \cdot \log_2 (1-p). \tag{3}$$

So distortion is a real valued parameter θ in the interval [0,1] with zero representing no distortion and one representing pure noise. It is the amount of entropy imposed on each bit of the plaintext sequence.

There is no way to estimate θ without coming to grips with the plaintext sequence $X_1, X_2, \cdots$; it must be modelled in some suitable way. To view it merely as a "nuisance parameter" is to render the task of estimating θ hopeless.

A tempting simple approach is unsatisfactory. If one assumes the elements of the plaintext sequence are iid Bernoulli random variables – an unrealistic assumption – one can obtain, asymptotically, an inequality for θ, but cannot obtain a consistent sequence of estimators. A contention of this paper is that there is a reasonable nonstochastic description of the plaintext sequence which does permit a consistent estimation of θ. Simulation studies, described in Section 3, evidence this with a real plaintext.

For a fixed positive integer value m, form a "grouped plaintext sequence" of "words" of length m:

$$x_1 = (X_1,X_2,\cdots,X_m),\ x_2 = (X_2,\cdots,X_{m+1}),\ x_3 = (X_3,\cdots,X_{m+2}),\ \cdots. \tag{4}$$

Likewise, form grouped sequences of length–m words using the Y and Z sequences: y_1, y_2, $\cdots$ and z_1, z_2, $\cdots$. The latter will be referred to as the "(grouped) *distorted* sequence."

Let G_m, with addition symbol "$\oplus$", denote the group of length–m vectors whose components are zeros and ones. Group addition means component–wise addition modulo 2. The (lower case) x's, y's and z's assume values in this group, and satisfy

$$z_n = x_n \oplus y_n,\ x_n = y_n \oplus z_n, \text{ and } \ y_n = x_n \oplus z_n. \tag{5}$$

The identity element is the vector of m zeros, $e = (0,0,\cdots,0)$. Below, $E_n(x)$ and $F_n(z)$ are functions on G_m, the empirical distributions of x_1, $\cdots$, x_n and z_1, $\cdots$, z_n, respectively. Notice that, for each fixed n, the function $\min\{E_n(x)\colon x \in G_m\}$ is nonincreasing in m.

We are now ready to describe a workable model for the plaintext sequence. Only one assumption is needed:

The Basic Assumption: *For some positive integer* m, $\min\{E_n(x)\colon x \in G_m\} \to 0$ *as* $n \to \infty$.

Clearly, when $\min\{E_n(x)\colon x \in G_m\} \to 0$ holds for some m, it holds for all larger m. The smallest such index m will be called the *complexity index* of the plaintext sequence.

We believe the basic assumption is reasonable. This is evidenced in Section 3 with the use of a Latin text that is well modelled with a complexity index of nine. We suspect that small complexity indices will adequately model many other complex structures that are subject to distortion.

For any two real-valued functions f and g defined on G_m, let "f*g" denote the convolution: $f*g(x) = \Sigma_{z \in G_m} f(z)g(x \oplus z)$, $x \in G_m$. Further, let $\gamma(x)$ denote the number of ones in the vector x, $x \in G_m$. Clearly, $0 \leq \gamma(x) \leq m$, and $\gamma(e) = 0$. Finally, let

$$P(y;p) := p^{\gamma(y)}(1-p)^{m-\gamma(y)}, \; y \in G_m, \tag{6}$$

which is the probability that $y_i = y$, $y \in G_m$, $i \geq 1$. The empirical distribution $F_n(z)$ for the distorted partial sequence $z_1, z_2, \cdots, z_n$ behaves, asymptotically, like the convolution $E_n*P(;p)(z)$. Specifically, the latter is the expected value of $F_n(z)$ (Lemma 1 below), and the difference $F_n(z) - E_n*P(;p)(z)$ goes to zero with probability one as $n \to \infty$, a consequence of Kolmogorov's strong law of large numbers (Lemma 2 below).

The empirical distribution $F_n(z)$, which describes a portion of the distorted text, is something that can be observed. The empirical distribution $E_n(x)$ for the undistorted text can not. If it could be observed, it would be an easy matter to obtain a consistent estimator of p and, hence, of θ. To see this, consider the function

$$Q(y;p) := (1 - 2p)^{-m}(-p)^{\gamma(y)}(1-p)^{m-\gamma(y)}, \; y \in G_m, \tag{7}$$

which is defined unless $p = \frac{1}{2}$. It can be checked (see Lemma 4 below), that the convolution $P(;p)*Q(;p)$ is the identity function $1_{\{e\}}$, i.e., $P(;p)*Q(;p)(x) = 1$ if $x = e$, and it is zero otherwise. This is extremely useful. For then, $E_n*P(;p)*Q(;p) = E_n$. It follows that $E_n(x)$ and the convolution $F_n*Q(;p)(x)$ are asymptotically equivalent. Specifically, the difference $F_n*Q(;p)(x) - E_n(x)$ converges to zero with probability one as $n \to \infty$, $x \in G_m$. Thus, *if* E_n *were known*, one should be able to obtain a consistent sequence of estimators of p by solving the equation $F_n*Q(;p)(x) = E_n(x)$ for p, using some suitable $x \in G_m$.

The assumption that $E_n(x)$ is known, upon which the previous paragraph builds, is untenable in practice. Nevertheless, one can show that $\{\theta_n = -p_n \cdot \log_2 p_n - (1-p_n) \cdot \log_2(1-p_n),\ n \geq 1\}$ is a consistent sequence of estimates for θ, where p_n is chosen to make $\min\{F_n{*}Q(;p_n)(x)\colon x \in G_m\} = 0$ (which is the limit of $\min\{E_n(x)\colon x \in G_m\}$ as $n \to \infty$), $n \geq 1$. The theory behind this is discussed in Section 2, and applications to various distorted plaintexts are discussed in Section 3.

2. Theory.

Here, the notation and concepts introduced in Section 1 are assumed. Our main result is:

Theorem. *Let* $p_n \in [0,\frac{1}{2}) \cup (\frac{1}{2},1]$ *be a root of the equation* $\min\{F_n{*}Q(;p)(x)\colon x \in G_m\} = 0$, *or let it be* $\frac{1}{2}$ *if no such root exists*, $n \geq 1$. *Then* $\theta_n = -p_n \cdot \log_2 p_n - (1-p_n) \cdot \log_2(1-p_n)$ *is uniquely defined, and, if the complexity index does not exceed* m, θ_n *is a strongly consistent estimator of the distortion index* θ. *That is,* $\theta_n \to \theta$ *with probability one as* $n \to \infty$.

The proof of this result depends on several lemmas, which we will discuss first.

Lemma 1. *The empirical distribution* F_n *for* $z_1, \cdots, z_n$ *has expectation* $\mathcal{E}F_n = E_n{*}P(;p)$.

Proof. The expectation of the indicator $1_{\{z\}}(z_i)$ is equal to

$$\Pr(z_i = z) = \Pr(x_i \oplus y_i = z) = \Pr(y_i = x_i \oplus z) = P(x_i \oplus z;p)$$

(see (5)). Thus, the expectation of $F_n(z)$ is:

$$\mathcal{E}F_n(z) = \frac{1}{n}\sum_{i=1}^{n} P(x_i \oplus z;p) = \frac{1}{n}\sum_{i=1}^{n} \Sigma_{x \in G_m} 1_{\{x\}}(x_i)(x \oplus z;p) = \Sigma_{x \in G_m}\{\frac{1}{n}\sum_{i=1}^{n} 1_{\{x\}}(x_i)\}P(x \oplus z;p) = E_n{*}P(z;p). \quad \square$$

Lemma 2. *The difference* $F_n(z) - \mathcal{E}F_n(z) \to 0$ *with probability one as* $n \to \infty$, $z \in G_m$.

Proof. Write $F_n(z) - \mathcal{E}F_n(z)$ as $n^{-1}\Sigma_{i=1}^{n} u_i$, where $u_i = 1_{\{z\}}(z_i) - P(x_i \oplus z;p)$. (See the proof of Lemma 1.) Because of the way the z_i's are defined (cf. (4)), u_r and u_s are dependent unless $|s - r| \geq m$. So split $n^{-1}\Sigma_{i=1}^{n} u_i$ up into m parts as $n^{-1}\Sigma_{j=1}^{m} v_{nj}$, where v_{nj} is the sum of the u_i with i of the form $j + mk$ for some nonnegative integer k, $1 \leq i \leq n$. It is easy to show, with Kolmogorov's strong law of large numbers (see page 165 of Stout (1974)), that $n^{-1}v_{nj} \to 0$ almost surely as $n \to \infty$, $1 \leq j \leq m$.

(Each u_i has mean zero and a variance bounded by one.) □

It will prove convenient, now, to extend the definition of $P(y;p)$, given in (6), to negative values of p; no modification of the formula is needed.

LEMMA 3. *For each real* p *and* q, $P(;p)^*P(;q) = P(;p+q-2pq)$.

PROOF. For fixed x and y in G_m, let j be the number of components of these vectors which are both one. It is easily seen that $j = (\gamma(x) + \gamma(y) - \gamma(x\oplus y))/2$. Also, for a fixed x and integer i, $0 \leq i \leq m$, the number of y in G_m for which $\gamma(y) = i$, and for which a particular value j is attained, is $\left[\begin{matrix}\gamma(x)\\ j\end{matrix}\right]\left[\begin{matrix}m\text{-}\gamma(x)\\ i-j\end{matrix}\right]$, (a simple "urn calculation"). It follows, for each fixed x in G_m, that

$$\begin{aligned}
&P(;p)^*P(;q)(x) = \Sigma_{y\in G_m} P(y;p)P(x\oplus y;q)\\
&= \Sigma_{y\in G_m} p^{\gamma(y)}(1-p)^{m-\gamma(y)}q^{\gamma(x\oplus y)}(1-q)^{m-\gamma(x\oplus y)}\\
&= \sum_{i=0}^{m} \Sigma_{\{y:\gamma(y)=i\}}\, p^i(1-p)^{m-i}q^{\gamma(x\oplus y)}(1-q)^{m-\gamma(x\oplus y)}\\
&= \sum_{i=0}^{m}\sum_{j=0}^{m}\left[\begin{matrix}\gamma(x)\\ j\end{matrix}\right]\left[\begin{matrix}m\text{-}\gamma(x)\\ i-j\end{matrix}\right] p^i(1-p)^{m-i}q^{\gamma(x)+i-2j}(1-q)^{m-\gamma(x)-i+2j}\\
&= \sum_{j=0}^{m}\left\{\left[\begin{matrix}\gamma(x)\\ j\end{matrix}\right](p(1-q))^j(q(1-p))^{\gamma(x)-j}\sum_{i=0}^{m}\left[\begin{matrix}m\text{-}\gamma(x)\\ i-j\end{matrix}\right](pq)^{i-j}((1-p)(1-q))^{m-\gamma(x)-(i-j)}\right\}\\
&= (p+q-2pq)^{\gamma(x)}(1-(p+q-2pq))^{m-\gamma(x)} = P(x;p+q-2pq). \qquad \square
\end{aligned}$$

With the simple observation $Q(;q) = P(;\frac{-q}{1-2q})$, $q \neq \frac{1}{2}$, Lemma 3 immediately yields:

LEMMA 4. *For* $q \neq \frac{1}{2}$, $Q(;q)^*P(;p) = P(;\frac{p-q}{1-2q})$. *Thus* $Q(;p)^*P(;p) = P(;0) = 1_{\{e\}}$ *when* $p \neq \frac{1}{2}$.

In the opposite direction, one can easily see that $P(;p) = Q(;\frac{-p}{1-2p})$ when $p \neq \frac{1}{2}$.

LEMMA 5. *For any nonconstant function* f *on* G_m, *the function* $h(p) := \min\{f^*Q(;p)(x): x \in G_m\}$, $p \in [0,\frac{1}{2})$, *is strictly decreasing and continuous, and its range is* $(-\infty,\min\{f(x): x \in G_m\}]$. (*If the domain of the function* h *is extended to the interval* $(\frac{1}{2},1]$, *it is easily seen that* $h(p) = h(1-p)$, $p \in (\frac{1}{2},1]$. *If* f *were constant on* G_m *with common value* c, *then* $h(p) = c$ *for all* p.)

PROOF. The continuity of h is inherited from the continuity of the functions $Q(;\cdot)(x)$, $x \in G_m$.

For fixed p and p′, $0 \leq p < p' < \frac{1}{2}$, let $q = \frac{p'-p}{1-2p} \in (0,\frac{1}{2})$, so that, according to Lemma 4, $Q(;p) = P(;q)^*Q(;p')$. Then, because the $P(z;q)$'s are nonnegative and add to unity,

$$\begin{aligned} f^*Q(;p)(x) &= P(;q)^*f^*Q(;p')(x) = \Sigma_{z\in G_m} P(z;q)\cdot f^*Q(;p')(x\oplus z) \\ &\geq \min\{f^*Q(;p')(x\oplus z): z \in G_m\} = h(p'), \end{aligned} \tag{8}$$

for each $x \in G_m$. So $h(p') \leq \min\{f^*Q(;p)(x): x \in G_m\} = h(p)$. Clearly, $h(p)$ is strictly greater than $h(p')$ if the inequality in (8) is always strict. If, on the contrary, it is an equality for some $x \in G_m$, then the function $f^*Q(;p')(\cdot)$ is constant, since the $P(z;q)$'s are, in fact, strictly positive. It would then follow that the function $f^*Q(;p)(\cdot)$ $(= P(;q)^*f^*Q(;p')(\cdot))$ is constant. Setting $p = 0$, leads to the contradiction that f is constant. Thus h is strictly decreasing on $[0,\frac{1}{2})$. It remains to show that $h(p) \to -\infty$ as $p \to \frac{1}{2}$. Suppose, to the contrary, that h is bounded below. Observe, that the sums of $f(x)$ and $f^*Q(;p)(x)$, $x \in G_m$, are equal (and finite) for each fixed $p \in [0,\frac{1}{2})$. From this, from the boundedness of h, and from the fact that G_m is a finite set, it easily follows that $\max\{f^*Q(;p)(x): x \in G_m\}$, $p \in [0,\frac{1}{2})$, is bounded *above*. Thus there must be a sequence $\{p_i\}$ converging up to $\frac{1}{2}$ on which $f^*Q(;\cdot)(x)$ converges to a finite limit, call it $g(x)$, $x \in G_m$. Then

$$f(x) = f^*Q(;p_i)^*P(;p_i)(x) \to g^*P(;\tfrac{1}{2})(x),$$

as $i \to \infty$, i.e., $f(x) = g^*P(;\frac{1}{2})(x)$ for all x. But $g^*P(;\frac{1}{2})(x)$ will be a constant function of x (follows immediately from the fact that $P(x,\frac{1}{2}) = 2^{-m}$ for all x). This says that f is a constant function of x, a contradiction. Thus, $h(p) \to -\infty$ as $p \to \frac{1}{2}$. □

LEMMA 6. *Let* f *be a nonnegative function on* G_m *whose values add to unity. Then*

$$\min\{f^*Q(;q)(x): x \in G_m\} \leq (1 - (2q)^m)^{-1}(\min\{f(x): x \in G_m\} - q^m),\ 0 \leq q < \tfrac{1}{2}.$$

PROOF. Observe that the function $h(x) := (1 - (2q)^m)^{-1}(P(x;q) - q^m)$, $x \in G_m$, is nonnegative and sums to unity for $q \in [0,\frac{1}{2})$. Thus for each $y \in G_m$,

$$\min\{f^*Q(;q)(x): x \in G_m\} = \min_{x\in G_m} \Sigma_{z\in G_m} Q(x\oplus z;q)f(z)$$

$$\leq \Sigma_{x \in G_m} h(x \oplus y) \Sigma_{z \in G_m} Q(x \oplus z;q) f(z)$$

$$= \Sigma_{z \in G_m} f(z) \{\Sigma_{x \in G_m} h(x \oplus y) Q(x \oplus z;q)\}$$

$$= (1 - (2q)^m)^{-1} \Sigma_{z \in G_m} f(z) \{1_{\{e\}}(y \oplus z) \ - q^m\}$$

$$= (1 - (2q)^m)^{-1} \{f(y) \ - q^m\}.$$

Since y is arbitrary, the desired conclusion follows. □

PROOF OF THE THEOREM. The heart of the proof is contained in Lemmas 5 and 6. According to Lemma 5, if the empirical distribution function F_n is not a constant, then the function $h_n(p) := \min\{F_n{*}Q(;p)(x)\colon x \in G_m\}$ is continuous and strictly decreasing on $[0,\frac{1}{2})$, and it has range $(-\infty,\min\{F_n(x)\colon x \in G_m\}]$. So $h_n(0) \geq 0$ and $h_n(p) \to -\infty$ as $p \to \frac{1}{2}$ So the equation $h_n(p) = 0$ has a unique root p' in the interval $[0,\frac{1}{2})$. This could be the root p_n referred to in the statement of the theorem. Alternatively, there is a second root at the point $1 - p'$, in the interval $(\frac{1}{2},1]$ (see the statement of Lemma 5), which could be the root p_n. Both possibilities give rise to the same unique value for θ_n. The case of a constant empirical function remains – when n is a multiple of 2^m (the cardinality of G_m): $F_n(x) = 2^{-m}$, $x \in G_m$. In such a case, there can be no *root* p_n (see the statement of Lemma 5), and, according to the statement of the theorem, p_n is set equal to $\frac{1}{2}$. Again, the value of θ_n is unique.

It remains to show that $\theta_n \to \theta$, with probability one, as $n \to \infty$. Let the true value of $p \in [0,1]$ be denoted by p_0. The task is to show that the only limit points of $\{p_n\}$ are p_0 and $1 - p_0$. Since $h_n(p) = h_n(1-p)$ for all $p \in [0,\frac{1}{2}) \cup (\frac{1}{2},1]$, there is no loss in generality in assuming that p_0 and the p_n are in the interval $[0,\frac{1}{2}]$. Then the task becomes that of showing $p_n \to p_0$ with probability one. For this, we shall need the fact, provided by Lemmas 1 and 2, that

$$F_n(x) = E_n{*}P(;p_0) + o(1), \text{ as } n \to \infty, \text{ uniform in } x, x \in G_m, \tag{9}$$

(uniformly since G_m is a finite set). There are two cases to consider: $p_0 \in [0,\frac{1}{2})$ and $p_0 = \frac{1}{2}$.

If $p_0 \in [0,\frac{1}{2})$, then $F_n{}^*Q(;p_0)(x) = E_n{}^*Q(;p_0)^*P(;p_0)(x) + o(1) = E_n(x) + o(1)$, as $n \to \infty$, which rules out the possibility that F_n is constant on G_m infinitely often. For, by assumption, the complexity index (defined in Section 1) does not exceed m, so that, according to the basic assumption (appearing in Section 1), $\min\{E_n(x): x \in G_m\} \to 0$ as $n \to \infty$. Thus $\min\{F_n{}^*Q(;p_0)(x): x \in G_m\} \to 0$. But, when F_n is constant on G_m, $F_n{}^*Q(;p_0)(x) = 2^{-m}$ *for all* $x \in G_m$. So F_n is constant on G_m only finitely often. These occurrences can be ignored, since we are only concerned about large n: For large n, $p_n \in [0,\frac{1}{2})$ and $h_n(p_n) = 0$. It remains to show for $p \in [0,\frac{1}{2})$ and $|p - p_0| \geq \delta$ that $h_n(p)$ cannot be zero when n is sufficiently large (depending on δ), $\delta > 0$.

Suppose $p_0 > 0$ and $\delta > 0$ is small enough that $p_0 - \delta > 0$. Then, for $p \in [0,p_0 - \delta]$, (9) and an application of Lemma 4 yield

$$\begin{aligned} h_n(p) &= \min\{F_n{}^*Q(;p)(x): x \in G_m\} \\ &= \min\{E_n{}^*P(;p_0)Q(;p)(x): x \in G_m\} + o(1) \\ &= \min\{E_n{}^*P(;\tfrac{p_0-p}{1-2p})(x): x \in G_m\} + o(1) \\ &\geq (\tfrac{p_0-p}{1-2p})^m + o(1) \geq (\delta/(1-2p_0))^m + o(1), \end{aligned} \tag{10}$$

which is strictly positive for all sufficiently large n. Thus $\liminf_{n\to\infty} p_n \geq p_0 - \delta$.

Likewise, for $p \in [p_0 + \delta,\frac{1}{2})$, (9) and an application of Lemmas 4 and 6 yield $P(;p_0)Q(;p) = Q(;\frac{p-p_0}{1-2p_0})$ and

$$\begin{aligned} h_n(p) &= \min\{F_n{}^*Q(;p)(x): x \in G_m\} \\ &= \min\{E_n{}^*P(;p_0)Q(;p)(x): x \in G_m\} + o(1) \\ &= \min\{E_n{}^*Q(;\tfrac{p-p_0}{1-2p_0})(x): x \in G_m\} + o(1) \\ &\leq (1 - 2\tfrac{p-p_0}{1-2p_0})^{-1}(\min\{E_n(x): x \in G_m\} - (\tfrac{p-p_0}{1-2p_0})^m) + o(1) \\ &\leq (1 - 2\tfrac{p-p_0}{1-2p_0})^{-1}(\{\min(E_n(x): x \in G_m\} - (\tfrac{\delta}{1-2p_0})^m) + o(1). \end{aligned} \tag{11}$$

Since $\min\{E_n(x): x \in G_m\} \to 0$ as $n \to \infty$, according to the basic assumption discussed in Section 1, the latter is strictly negative for sufficiently large n. Thus $\limsup_{n\to\infty} p_n \leq p_0 + \delta$. This complete the argument for $p_0 \in [0,\frac{1}{2})$.

Finally, suppose $p_0 = \frac{1}{2}$. Here, we do not know whether or not F_n is constant on G_m infinitely often; it does not matter. When it is, $p_n = \frac{1}{2}$ (according to the statement of the theorem), and $p_n = p_0$. When it is not, the argument shown in (10) will suffice: For no $\delta > 0$, can p_n be less than $\frac{1}{2} - \delta$ when n is sufficiently large. □

3. **Applications.** Because computer files of natural language texts are readily available, it is convenient to apply our methodology to deliberately distorted natural language texts. We shall work with Virgil's Aeneid, in the original Latin text. The approach is to convert each of the 26 letters of the Roman alphabet to five binary bits, in order, with decimal equivalents 0 to 25. The remaining six 5–bit numbers, with decimal equivalents 26 to 31, are used for five punctuation marks and the delimiter between words (a "space"). Each bit is changed with probability $p \in [0,\frac{1}{2}]$, where p is chosen to correspond to a prescribed distortion parameter value θ. The Aeneid requires about 2.1 million bits. It is well modelled with a complexity index of nine; one of the 512 possible binary bit sequences of length nine *never* occurs, while *all* of the 256 binary bit sequences of length eight do occur at least once. (The index may depend on the way we code the Roman alphabet; this issue has not been addressed.)

Using the methods described in Sections 1 and 2, we obtained, as shown in Table 2 below, the estimates of θ for $\theta = .05, .10, \cdots, 1.00$.

Actual Distortion	Estimated Distortion	Actual Distortion	Estimated Distortion
.05	.0538	.55	.5264
.10	.1093	.60	.5682
.15	.1444	.65	.5937
.20	.1961	.70	.6124
.25	.2532	.75	.6750
.30	.2960	.80	.6787
.35	.3302	.85	.7260
.40	.3817	.90	.7261
.45	.4439	.95	.7335
.50	.4565	1.00	.7483

Table 2: Estimated Distortion for Various Distortion Levels (n = 2,000,000)

These estimates use the empirical distribution function F_n for two–million overlapping 9–bit distorted sequences, intentional distortions of the plaintext. The estimates for $\theta \leq .5$, appearing in the second column, are quite respectable. The estimates for $\theta \geq .55$, appearing in the fourth column, show a pronounced negative bias. This occurrence of a negative bias seems to be due to fluctuations brought about by the randomization of bits. As one might expect, the problem of bias increases if one uses less of the distorted text to estimate θ, and it occurs at smaller values of θ. (See Table 3 below.)

Bias is a greater problem than variance. To illustrate this persuasively, the two million 9–bit sequences used to estimate a particular value of θ for Table 2 were split up into 20 equal parts and used to make 20 (essentially independent) estimates of θ. Sample results are shown in Table 3 for ten different values of θ.

Actual Distortion	Bias (estimated)	Standard Dev. (estimated)
.10	−.0157	.0162
.20	−.0408	.0218
.30	−.0583	.0255
.40	−.0859	.0212
.50	−.1123	.0311
.60	−.1539	.0312
.70	−.2127	.0293
.80	−.2718	.0213
.90	−.3589	.0258
1.00	−.4536	.0221

Table 3: Bias and Standard Deviation of the Estimator
(20 cases with n = 100,000)

As the table shows, the standard deviation of the estimator of θ stays fairly constant and relatively small while the bias grows increasingly negative – to a disturbing size – as θ approaches one.

The relative smallness of the standard deviation offers some hope that accurate estimates of bias can be found. When used, these should result in accurate, moderate–sample–size, estimates of distortion. Besides helping with the bias, a "second order theory", would probably point the way to a useful central limit theorem.

It is perhaps worth emphasizing that huge values of n can be expected with distorted visual images, and probably in most contexts. I.e., large sets of binary bits can be anticipated. So accurately estimating distortion should be feasible for a wide variety of settings.

A question of considerable importance remains: How does one determine the complexity index? This index must either be inferred from experience or be inferred from the distorted image itself. We believe the latter may be possible.

The reason for such optimism is that (i) the estimators θ_n converge to the correct limit θ when the complexity index does *not* exceed m (see the theorem of Section 2), and (ii) the difference $\theta_n - \theta$ has to have a strictly positive limit point when the complexity index exceeds m (a consequence mainly of Lemma 5). Negative limit points are impossible. Moreover, a limit point at zero is ruled out unless $\min\{E_n(x): x \in G_m\}$, $n \geq 1$, has a limit point at zero. (The latter won't converge to zero when the index exceeds m.) By experimenting with several values of m, it should be possible to distinguish between cases (i) and (ii), when n is large, and correctly estimate the true value of the complexity index. This agenda should be easier to implement when the issue of bias, discussed above, is better understood.

Acknowledgement

The author wishes to acknowledge, with appreciation, various conversations with his colleague Douglas Kelly that were helpful with the formulation of some of the ideas of this paper.

References

Billingsley, Patrick (1965). *Ergodic Theory and Information*, Wiley, New York.

Stout, William (1974). *Almost Sure Convergence*, Academic Press, New York.