From absolute distinguishability to positive distinguishability

Zvika Brakerski and Oded Goldreich*
Department of Computer Science
Weizmann Institute of Science, Rehovot, Israel.

March 9, 2009

Abstract

We study methods of converting algorithms that distinguish pairs of distributions with a gap that has an absolute value that is noticeable into corresponding algorithms in which the gap is always positive. Our focus is on designing algorithms that, in addition to the tested string, obtain a fixed number of samples from each distribution. Needless to say, such algorithms can not provide a very reliable guess for the sign of the original distinguishability gap, still we show that even guesses that are noticeably better than random are useful in this setting.

Keywords: Computational Indistinguishability, Statistical Indistinguishability.

1 The problem and its solutions

This note addresses a generic technical problem that arises in the context of trying to establish the computational indistinguishability of certain pairs of probability ensembles. The problem refers to the fact that computational (an also statistical) indistinguishability are defined in terms of the absolute difference between probabilities, whereas it is typically easier to manipulate the difference itself. Thus, we seek a method of converting a noticeable absolute difference into a noticeable difference; that is, the difference itself (rather than its absolute value) should be positive.

1.1 A motivational example

Many security definitions are formulated by referring to two pairs of *probability ensembles that are indexed by strings*, and requiring that these pairs of probability ensembles are computationally indistinguishable (see, e.g., the definitions of computational zero-knowledge [1, Sec. 4.3.1.2] and secure two-party computation [2, Sec. 7.2]). Such a probability ensemble $\{Z_\alpha\}_{\alpha\in S}$ consists of (an infinite number of) "random variables" Z_α's, which are each distributed over some finite set (related to its index, α). Two such ensembles, $\{X_\alpha\}_{\alpha\in S}$ and $\{Y_\alpha\}_{\alpha\in S}$, are said to be computationally indistinguishable if for every probabilistic polynomial-time algorithm D it holds that

$$g_D(\alpha) \stackrel{\text{def}}{=} |\Pr[D(\alpha,X_\alpha)=1] \;-\; \Pr[D(\alpha,Y_\alpha)=1]| \tag{1}$$

is negligible as a function of $|\alpha|$ (i.e., for every positive polynomial p and all sufficiently long α's the value of $g_D(\alpha)$ is upper bounded by $1/p(|\alpha|)$).

*Partially supported by the Israel Science Foundation (grant No. 1041/08).

The aforementioned formulation mandates that the value of $g_D(\alpha)$ is small for every $\alpha \in S$. A weaker requirement, which suffices in practice, is that it is infeasible to find $\alpha \in S$ for which the value of $g_D(\alpha)$ is not small. This requirement may be formulated as mandating that for every probabilistic polynomial-time algorithm F, representing a potential finder that given 1^n outputs an n-bit long string $\alpha \in S$, the expected value of $g_D(\alpha)$ (when defined as in Eq. (1)) is negligible (as a function of n); that is, $E[g_D(F(1^n))]$ is negligible in n. This condition means that

$$\sum_\alpha \Pr[F(1^n)=\alpha] \cdot |\Pr[D(\alpha,X_\alpha)=1] \;-\; \Pr[D(\alpha,Y_\alpha)=1]| \tag{2}$$

is negligible as a function of n.

When trying to establish a condition as in Eq. (2) it is often easier to establish a corresponding condition in which the absolute value operator is dropped. Indeed, suppose that for every F and D as above it holds that

$$\sum_\alpha \Pr[F(1^n)=\alpha] \cdot (\Pr[D(\alpha,X_\alpha)=1] \;-\; \Pr[D(\alpha,Y_\alpha)=1]) \tag{3}$$

is negligible (as a function of n). Can we infer that Eq. (2) holds too?

In the case that both ensembles are polynomial-time sampleable, a positive answer is implicit in many works. Basically, given a probabilistic polynomial-time algorithm D such that Eq. (2) is not negligible, one derives a probabilistic polynomial-time algorithm D' such that Eq. (3) is not negligible by estimating the difference between $\Pr[D(\alpha,X_\alpha)=1]$ and $\Pr[D(\alpha,Y_\alpha)=1]$ and flipping D's output if the estimated difference is negative. Thus, the construction of D' depends also on g_D (which determines the adequate level of approximation).

1.2 A generic problem and one solution

The generic problem we face is converting an algorithm D that distinguishes X_α and Y_α (i.e., $|\Pr[D(\alpha,X_\alpha)=1] - \Pr[D(\alpha,Y_\alpha)=1]|$ is noticeable) into an algorithm D' that on input (α,X_α) outputs 1 with probability that is noticeably higher than $\Pr[D(\alpha,Y_\alpha)=1]$. We stress that we wish this transformation to hold for every α, and it may be that for some α's the difference $\Pr[D(\alpha,X_\alpha)=1] - \Pr[D(\alpha,Y_\alpha)=1]$ is positive whereas for other α's the difference is negative. Clearly, D' must know something about X_α and Y_α in order for this to be possible, and indeed we provide D' with samples taken from X_α and Y_α (or, actually, with algorithms for sampling these distributions).

Thus, the problem we face is actually the following one. We are given a probabilistic polynomial-time algorithm D and sampling algorithms for two ensembles, $\{X_\alpha\}_{\alpha\in S}$ and $\{Y_\alpha\}_{\alpha\in S}$ (i.e., probabilistic polynomial-time algorithms X and Y such that on any input α it holds that $X_\alpha \equiv X(\alpha)$ and $Y_\alpha \equiv Y(\alpha)$). Our task is to construct a probabilistic polynomial-time algorithm D' such that for some function $\rho : (0,1] \to (0,1]$ it holds that

$$\Pr[D'(\alpha,X_\alpha)=1] \;-\; \Pr[D'(\alpha,Y_\alpha)=1] \;\geq\; \rho\,(|\Pr[D(\alpha,X_\alpha)=1] \;-\; \Pr[D(\alpha,Y_\alpha)=1]|)\,. \tag{4}$$

We stress that the r.h.s of Eq. (4) refers to the *absolute* difference between two probabilities, whereas the l.h.s refers to a corresponding difference that is not taken in absolute value and yet is required to be positive (whenever the former difference is positive).

We seek a universal transformation of D into D', whereas this transformation may use a predetermined number of auxiliary samples of the two distributions. That is, the resulting algorithm D' is given as input a single sample that is drawn from one of two distributions (i.e., either from

X_α or from Y_α), but in addition it can obtain (a predetermined number of) samples from each of the two distributions. Like D, algorithm D' should distinguish the two cases (which correspond to the source of its input). We stress that we wish the complexity of D' (and specifically the number of auxiliary samples it obtains) to be independent of $g_D(\alpha)$. We note that such a transformation (of D into D') may be useful also in other settings.

A simple transformation. One solution to this problem is to let D' estimate the sign of $\Pr[D(\alpha, X_\alpha) = 1] - \Pr[D(\alpha, Y_\alpha) = 1]$ by using a single sample of X_α and a single sample of Y_α. (Although this estimate is quite poor, it can be shown to suffice.) Specifically, on input (α and) z (where z is taken from either X_α or Y_α), algorithm D' proceeds as follows:

1. Ignoring its ("main") input (i.e., z), algorithm D' generates a single sample x of X_α and a single sample y of Y_α, and computes $\sigma \leftarrow D(\alpha, x)$ and $\tau \leftarrow D(\alpha, y)$;

2. If $\sigma > \tau$ then D' invokes D on its inputs and outputs $D(\alpha, z)$.

   If $\sigma < \tau$ then D' outputs $1 - D(\alpha, z)$.

   Otherwise (i.e., $\sigma = \tau$), algorithm D' outputs the outcome of a fair coin toss.

Indeed, we have assumed here (without loss of generality) that D always outputs a Boolean value. Intuitively, $\sigma - \tau$ provides a probabilistic guess of the sign of $\Pr[D(\alpha, X_\alpha)=1] - \Pr[D(\alpha, Y_\alpha)=1]$, and using this guess in the obvious manner yields the desired result. For the actual analysis of the performance of D', we consider an algorithm D'', which may output any number in $[0, 1]$, such that

$$D''(\alpha, z) \stackrel{\text{def}}{=} \frac{1}{2} \cdot \left(1 + \texttt{sign}(D(\alpha, X_\alpha) - D(\alpha, Y_\alpha)) \cdot (-1)^{D(\alpha,z)+1}\right), \tag{5}$$

where $\texttt{sign}(r) = 1$ if $r > 0$ (resp., $\texttt{sign}(r) = -1$ if $r < 0$), and $\texttt{sign}(0) = 0$. Recall that in Step 2 of $D'(\alpha, z)$, the output is set to $D(\alpha, z)$ if $\sigma > \tau$, to $1 - D(\alpha, z)$ if $\sigma < \tau$, and is random if $\sigma = \tau$. Using $D(\alpha, z) \in \{0, 1\}$ and assuming $\sigma \neq \tau$, the output of $D'(\alpha, z)$ can be written as $(1+\texttt{sign}(\sigma-\tau)\cdot(-1)^{D(\alpha,z)+1})/2$. Thus, $D'(\alpha, z)$ outputs 1 with probability $D''(\alpha, z)$, and it suffices to evaluate

$$\mathrm{E}[D''(\alpha, X_\alpha)] - \mathrm{E}[D''(\alpha, Y_\alpha)] \;=\; \Pr[D'(\alpha, X_\alpha)=1] - \Pr[D'(\alpha, Y_\alpha)=1]. \tag{6}$$

Denoting $p = \Pr[D(\alpha, X_\alpha)=1]$ and $q = \Pr[D(\alpha, Y_\alpha)=1]$ (and using X'_α and Y'_α to denote independent copies of X_α and Y_α), we evaluate Eq. (6) as follows.

$$\begin{aligned}
g_{D''}(\alpha) &\stackrel{\text{def}}{=} \mathrm{E}[D''(\alpha, X_\alpha)] - \mathrm{E}[D''(\alpha, Y_\alpha)] \\
&= \frac{1}{2} \cdot \mathrm{E}\left[1 + \texttt{sign}(D(\alpha, X'_\alpha) - D(\alpha, Y'_\alpha)) \cdot (-1)^{D(\alpha,X_\alpha)+1}\right] \\
&\quad -\frac{1}{2} \cdot \mathrm{E}\left[1 + \texttt{sign}(D(\alpha, X'_\alpha) - D(\alpha, Y'_\alpha)) \cdot (-1)^{D(\alpha,Y_\alpha)+1}\right] \\
&= \frac{1}{2} \cdot \mathrm{E}\left[\texttt{sign}(D(\alpha, X'_\alpha) - D(\alpha, Y'_\alpha))\right] \cdot \mathrm{E}\left[(-1)^{D(\alpha,X_\alpha)+1} - (-1)^{D(\alpha,Y_\alpha)+1}\right]
\end{aligned}$$

Using $\mathrm{E}[(-1)^{D(\alpha,X_\alpha)+1}] = p - (1-p) = 2p - 1$ and $\mathrm{E}[(-1)^{D(\alpha,Y_\alpha)+1}] = 2q - 1$, we get

$$\begin{aligned}
g_{D''}(\alpha) &= (p-q) \cdot \mathrm{E}\left[\texttt{sign}(D(\alpha, X_\alpha) - D(\alpha, Y_\alpha))\right] \\
&= (p-q) \cdot (\Pr[D(\alpha, X_\alpha) > D(\alpha, Y_\alpha)] - \Pr[D(\alpha, X_\alpha) < D(\alpha, Y_\alpha)]) \\
&= (p-q) \cdot (p \cdot (1-q) - (1-p) \cdot q)
\end{aligned}$$

which equals $(p-q)^2$.

1.3 Other transformations

Two natural questions arise:

1. Is the foregoing construction of D' optimal (with respect to all constructions that use a single auxiliary sample from each of the two distributions)?
2. Can we do better if we obtain k auxiliary samples from each of the two distributions (rather than one)? How good can such a procedure be?

Before answering these questions we note that no procedure (which is given a single test sample from an unknown distribution) can outperform the variation distance, that is, $|p-q|$. We answer the above questions as follows.

Main Result (informal). *For every $k \geq 1$, the best construction that uses k auxiliary samples from each of the two distributions rules analogously to Eq. (5), when applying the sign function to the difference between the average value of D in the two cases. Such a procedure yields a gap of at least the minimum of $\Omega(\sqrt{k}) \cdot (p-q)^2$ and $(1-\epsilon_{p,q}(k)) \cdot |p-q|$, where $\epsilon_{p,q}(k) = \exp(-\Omega((p-q)^2 \cdot k))$.*

We stress that the above result holds both in the computational setting and in the information theoretic setting.

2 The actual treatment

Let X and Y be 0-1 random variables (representing $D(\alpha, X_\alpha)$ and $D(\alpha, Y_\alpha)$, respectively), and let X_i's (resp., Y_i's) be independent copies of X (resp., Y) representing additional samples available to us. We seek a randomized process $\Pi : \{0,1\}^{2k+1} \to \{0,1\}$ such that

$$\mathrm{E}[\Pi(X_1, ..., X_k, Y_1, ..., Y_k, X)] - \mathrm{E}[\Pi(X_1, ..., X_k, Y_1, ..., Y_k, Y)] \tag{7}$$

is maximized (as a function $\delta = |\mathrm{E}[X] - \mathrm{E}[Y]|$, when maximizing over all possible 0-1 random variables X and Y that are at statistical distance δ). Indeed, the probability that $\Pi(a_1, ..., a_k, b_1, ..., b_k, c) = 1$ is determined by a function $f : \{0,1\}^{2k+1} \to [0,1]$ such that

$$\Pr[\Pi(a_1, ..., a_k, b_1, ..., b_k, c) = 1] = f(a_1, ..., a_k, b_1, ..., b_k, c)$$

Thus, it suffices to seek such a function f that maximizes

$$\mathrm{E}[f(X_1, ..., X_k, Y_1, ..., Y_k, X)] - \mathrm{E}[f(X_1, ..., X_k, Y_1, ..., Y_k, Y)] \tag{8}$$

(as a function $\delta = |\mathrm{E}[X] - \mathrm{E}[Y]|$).

Let us formally define a more general optimization problem. For a function $f : \{0,1\}^{2k+1} \to [0,1]$ and a pair $(p,q) \in [0,1]$, we denote by $\mathcal{V}_{(p,q)}(f)$ the value of Eq. (8), when X and Y satisfy $p = \mathrm{E}[X]$ and $q = \mathrm{E}[Y]$. Now, for any (possibly infinite) set (or class) of pairs in $[0,1]$, denoted $\mathcal{C}$, and any function $f : \{0,1\}^{2k+1} \to [0,1]$, we denote $\mathcal{V}_{\mathcal{C}}(f) \stackrel{\text{def}}{=} \min_{(p,q)\in\mathcal{C}}\{\mathcal{V}_{(p,q)}(f)\}$. We seek a function f for which $\mathcal{V}_{\mathcal{C}}(f)$ is maximal.

Overview. First, we will show that, without loss of generality, the function $f(x_1, ..., x_k, y_1, ..., y_k, z)$ may only depend on $s \stackrel{\text{def}}{=} \sum_{i\in[k]} x_i$, $t \stackrel{\text{def}}{=} \sum_{i\in[k]} y_i$ and z, and furthermore that it can take a specific canonical form (see Section 2.1). Next, in Section 2.2, we will show that in all natural cases (i.e., for "symmertic" classes) the canonical form can be further simplified to depend only on $\mathtt{sign}(s-t)$ and z. Actually, this will yield a single optimal function. Lastly, in Section 2.3, we will analyze the performance of this function.

2.1 Canonical functions

We will first show that it suffices to consider functions f of the form

$$f(a_1, ..., a_k, b_1, ..., b_k, c) = \frac{1 + g\left(\sum_{i\in[k]} a_i\, , \sum_{i\in[k]} b_i\right) \cdot (-1)^c}{2} \tag{9}$$

where $g : \mathbb{N}^2 \to [-1, +1]$. We call such an f canonical. Note that the normalization (i.e., shifting by 1 and dividing by 2) is used to map $[-1, +1]$ to $[0, 1]$. (Note that an additive shift of f leaves the value of Eq. (8) intact, whereas multiplying f by any factor has the same effect on the value of Eq. (8).)

Definition 2.1 (dominating strategies) *We say that f' dominates f (w.r.t $\mathcal{C}$) if for every $(p, q) \in \mathcal{C}$ it holds that $\mathcal{V}_{(p,q)}(f') \geq \mathcal{V}_{(p,q)}(f)$.*

Proposition 2.2 (strong optimality): *For every $\mathcal{C}$ and every $f : \{0, 1\}^{2k+1} \to [-1, +1]$ there exists a canonical function that dominates f.*

Proof: Given any function f, we consider the function f' such that for every $a, b \in \{0, 1, ..., k\}$ and $c \in \{0, 1\}$, the value $f'(a, b, c)$ equals the average of $f(a_1, ..., a_k, b_1, ..., b_k, c)$ taken over all $(a_1, ..., a_k), (b_1, ..., b_k) \in \{0, 1\}^k$ that satisfy $\sum_{i\in[k]} a_i = a$ and $\sum_{i\in[k]} b_i = b$. Then, for every (p, q), we have $\mathcal{V}_{(p,q)}(f') = \mathcal{V}_{(p,q)}(f)$. Note that the value of f' at any (a, b) and $c \in \{0, 1\}$ can be written as

$$\begin{aligned} & \frac{1 + (-1)^c}{2} \cdot f'(a, b, 0) + \frac{1 - (-1)^c}{2} \cdot f'(a, b, 1) \\ = \;& \frac{1}{2} \cdot (f'(a, b, 0) + f'(a, b, 1)) + \frac{(-1)^c}{2} \cdot (f'(a, b, 0) - f'(a, b, 1)) \\ = \;& g_0(a, b) + g_1(a, b) \cdot (-1)^c \end{aligned}$$

where $g_0(a, b) = (f'(a, b, 0) + f'(a, b, 1))/2$ and $g_1(a, b) = (f'(a, b, 0) - f'(a, b, 1))/2$. Note that $g_1(a, b) \in [-0.5, +0.5]$ and that replacing $g_0(a, b)$ by 0.5 does not change the value of $\mathcal{V}_{(p,q)}(f')$. Thus, setting $f''(a, b, c) = (1 + 2g_1(a, b) \cdot (-1)^c)/2$, we obtain a canonical function that dominates f (because $\mathcal{V}_{(p,q)}(f'') = \mathcal{V}_{(p,q)}(f') = \mathcal{V}_{(p,q)}(f)$). ■

Conclusion and Notation. At this point we can limit our search for good functions (i.e., functions that maximize Eq. (8)) to canonical functions. That is, for every function $g : \mathbb{N}^2 \times \{0, 1\} \to [-1, +1]$ and every $k \in \mathbb{N}$, we define $f_g^{(k)}$ as in Eq. (9), and consider the value $\mathcal{V}_{(p,q)}(f_g^{(k)})$. To estimate $\mathcal{V}_{(p,q)}(f_g^{(k)})$, we let X and Y be 0-1 random variables with $\mathrm{E}[X] = p$ and $\mathrm{E}[Y] = q$ and get

$$\mathcal{V}_{(p,q)}(f_g^{(k)}) \;=\; \frac{1}{2} \cdot \mathrm{E}\left[g\left(\sum_{i\in[k]} X_i\, , \sum_{i\in[k]} Y_i\right) \cdot (-1)^X\right] - \frac{1}{2} \cdot \mathrm{E}\left[g\left(\sum_{i\in[k]} X_i\, , \sum_{i\in[k]} Y_i\right) \cdot (-1)^Y\right] \tag{10}$$

Using the independence of X, Y and the X_i's and Y_i's, we rewrite Eq. (10) as

$$\mathcal{V}_{(p,q)}(f_g^{(k)}) \;=\; \frac{1}{2} \cdot \mathrm{E}\left[g\left(\sum_{i\in[k]} X_i\, , \sum_{i\in[k]} Y_i\right)\right] \cdot \mathrm{E}\left[(-1)^X - (-1)^Y\right]. \tag{11}$$

Recalling that $\mathrm{E}[(-1)^X] = (1-p)-p = 1-2p$ and $\mathrm{E}[(-1)^Y] = 1-2q$, we get $\mathrm{E}[(-1)^X - (-1)^Y] = 2(q-p)$ and so

$$\mathcal{V}_{(p,q)}(f_g^{(k)}) = (q-p) \cdot \mathrm{E}[g(X',Y')], \tag{12}$$

where $X' = \sum_{i\in[k]} X_i$ and $Y' = \sum_{i\in[k]} Y_i$. Denoting $B(p,i,k) = \binom{k}{i} \cdot p^i \cdot (1-p)^{k-i}$, we get

$$\mathcal{V}_{(p,q)}(f_g^{(k)}) = (q-p) \cdot \sum_{i,j\in\{0,1,\dots,k\}} B(p,i,k) \cdot B(q,j,k) \cdot g(i,j) \tag{13}$$

2.2 Symmetric classes

We focus on symmetric classes of pairs, where $\mathcal{C}$ is symmetric if for every $(p,q) \in \mathcal{C}$ it also holds that $(q,p) \in \mathcal{C}$. In contrast, if $\mathcal{C}$ contains only pairs (p,q) such that $p > q$, then we may set $k = 0$ and use the identity function (because $\mathrm{E}[X] - \mathrm{E}[Y] = p - q = \texttt{StatDiff}(X,Y)$). We show that, for symmetric classes, the "sign of the difference" function (i.e., $\texttt{sd}(a,b) = \texttt{sign}(b-a) \in \{-1,0,+1\}$) is optimal as a function g.

Proposition 2.3 (optimality): *For every symmetric $\mathcal{C}$ and every $k \in \mathbb{N}$ and $g : \mathbb{N}^2 \to [-1,+1]$, it holds that $\mathcal{V}_{\mathcal{C}}(f_{\texttt{sd}}^{(k)}) \geq \mathcal{V}_{\mathcal{C}}(f_g^{(k)})$, where $\texttt{sd}(a,b) = \texttt{sign}(b-a)$.*

Recall that $\texttt{sign}(d) = -1$ if $d < 0$ (resp., $\texttt{sign}(d) = 1$ if $d > 0$), and $\texttt{sign}(0) = 0$.

Proof: Let $(p,q) \in \mathcal{C}$ be such that $\mathcal{V}_{(p,q)}(f_{\texttt{sd}}^{(k)}) = \mathcal{V}_{\mathcal{C}}(f_{\texttt{sd}}^{(k)})$. Then, $\mathcal{V}_{\mathcal{C}}(f_g^{(k)}) \leq (\mathcal{V}_{(p,q)}(f_g^{(k)}) + \mathcal{V}_{(q,p)}(f_g^{(k)}))/2$ (by definition of $\mathcal{V}_{\mathcal{C}}(f_g^{(k)})$ and the fact that $(q,p) \in \mathcal{C}$ [which follows by the symmetry of $\mathcal{C}$]), whereas $\mathcal{V}_{\mathcal{C}}(f_{\texttt{sd}}^{(k)}) \geq \mathcal{V}_{(p,q)}(f_{\texttt{sd}}^{(k)})$ (by the choice of $(p,q) \in \mathcal{C}$). Also note that $\mathcal{V}_{(p,q)}(f_{\texttt{sd}}^{(k)}) = \mathcal{V}_{(q,p)}(f_{\texttt{sd}}^{(k)})$ (by the invariance of the function $f_{\texttt{sd}}^{(k)}$ under of this switch, as seen in Eq. (12)). Thus, it suffices to show that

$$\mathcal{V}_{(p,q)}(f_{\texttt{sd}}^{(k)}) + \mathcal{V}_{(q,p)}(f_{\texttt{sd}}^{(k)}) \geq \mathcal{V}_{(p,q)}(f_g^{(k)}) + \mathcal{V}_{(q,p)}(f_g^{(k)}). \tag{14}$$

For every $a,b \in \{0,1,...,k\}$, we shall show that replacing $g(a,b)$ by $\texttt{sign}(b-a)$ may only increase $\mathcal{V}_{(p,q)}(f_g^{(k)}) + \mathcal{V}_{(q,p)}(f_g^{(k)})$. Let us start by recalling Eq. (13), which yields

$$\begin{aligned}
\mathcal{V}_{(p,q)}(f_g^{(k)}) + \mathcal{V}_{(q,p)}(f_g^{(k)}) &= (q-p) \cdot \sum_{i,j\in\{0,1,\dots,k\}} B(p,i,k)B(q,j,k) \cdot g(i,j) \\
&\quad +(p-q) \cdot \sum_{i,j\in\{0,1,\dots,k\}} B(q,i,k)B(p,j,k) \cdot g(i,j) \\
&= (q-p) \cdot \sum_{i,j\in\{0,1,\dots,k\}} [B(p,i,k)B(q,j,k) - B(q,i,k)B(p,j,k)] \cdot g(i,j).
\end{aligned}$$

Clearly, for $i = j$ we have $B(p,i,k)B(q,j,k) = B(q,i,k)B(p,j,k)$. For $i < j$ (resp., $j < i$), it holds that $B(p,i,k)B(q,j,k) > B(q,i,k)B(p,j,k)$ if and only if $p < q$ (resp., $q < p$). The latter claim seems self-evident, yet we provide a detailed proof next (for the case $p,q \in (0,1)$).

$$\begin{aligned}
B(p,i,k)B(q,j,k) &= \binom{k}{i} \cdot p^i \cdot (1-p)^{k-i} \cdot \binom{k}{j} \cdot q^j \cdot (1-q)^{k-j} \\
&= \binom{k}{i} \cdot (1-p)^k \cdot \binom{k}{j} \cdot (1-q)^k \cdot (p/(1-p))^i \cdot (q/(1-q))^j
\end{aligned}$$

Thus, $\frac{B(p,i,k)B(q,j,k)}{B(q,i,k)B(p,j,k)}$ equals

$$\frac{(p/(1-p))^i \cdot (q/(1-q))^j}{(q/(1-q))^i \cdot (p/(1-p))^j} = \frac{(q/(1-q))^{j-i}}{(p/(1-p))^{j-i}}$$

Note that we have $p < q$ iff $(p/(1-p)) < (q/(1-q))$, and so $p < q$ iff $(p/(1-p))^{j-i} < (q/(1-q))^{j-i}$. It follows that $p < q$ iff $B(p,i,k)B(q,j,k) > B(q,i,k)B(p,j,k)$.

Recall that for $i < j$, it holds that $B(p,i,k)B(q,j,k) - B(q,i,k)B(p,j,k) > 0$ if and only if $q > p$. Thus, in this case, we maximize

$$(q-p) \cdot [B(p,i,k)B(q,j,k) - B(q,i,k)B(p,j,k)] \cdot g(i,j) \tag{15}$$

by setting $g(i,j) = 1$ (because the first two factors have the same sign). Similarly, for $j > i$, it holds that $B(p,i,k)B(q,j,k) - B(q,i,k)B(p,j,k) > 0$ if and only if $q < p$, and so the maximization requires $g(i,j) = -1$. Indeed, for $i = j$, any setting of $g(i,j)$ will do. Thus, an optimal setting of $g(i,j)$ is $\mathtt{sign}(j-i)$, which equals $\mathtt{sd}(i,j)$. The claim follows. ■

2.3 The performance of the function $f_{\mathtt{sd}}^{(k)}$

We now turn to evaluating the performance of the optimal function; that is, we evaluate $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$. Recall that

$$\begin{aligned}\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) &= (q-p) \cdot \sum_{i,j \in \{0,1,\ldots,k\}} B(p,i,k)B(q,j,k) \cdot \mathtt{sd}(i,j) \\ &= (p-q) \cdot \sum_{i,j \in \{0,1,\ldots,k\}} B(p,i,k)B(q,j,k) \cdot \mathtt{sign}(i-j)\end{aligned}$$

which yields $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) = (p-q) \cdot v_{p,q}$, where

$$v_{p,q} \stackrel{\text{def}}{=} \mathrm{E}\left[\mathtt{sign}\left(\sum_{i\in[k]} X_i - \sum_{i\in[k]} Y_i\right)\right] \tag{16}$$

such that the X_i's (resp., Y_i's) are 0-1 i.i.d with expectation p (resp., q). Letting $T_i = X_i - Y_i$, we rewrite Eq. (16) as $\mathrm{E}[\mathtt{sign}(\sum_{i\in[k]} T_i)]$, which equals

$$\Pr\left[\sum_{i\in[k]} T_i > 0\right] - \Pr\left[\sum_{i\in[k]} T_i < 0\right]. \tag{17}$$

Note that $\mathrm{E}[T_i] = p - q$ and $\mathrm{Var}[T_i] = p(1-p) + q(1-q)$.

The cases of $k = 1$ and $k = 2$. For small k, we can write explicit expressions for Eq. (17); for example, for $k = 1$ Eq. (17) yields $\Pr[T_1 > 0] - \Pr[T_1 < 0] = p(1-q) - q(1-p) = p - q$, and so $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(1)}) = (p-q)^2$. For $k = 2$, we have

$$\begin{aligned}\Pr[T_1 + T_2 > 0] - \Pr[T_1 + T_2 < 0] &= \Pr[T_1 + T_2 = 2] + 2\Pr[T_1 = 1 \wedge T_2 = 0] \\ &\quad - (\Pr[T_1 + T_2 = -2] + 2\Pr[T_1 = -1 \wedge T_2 = 0]) \\ &= p^2(1-q)^2 + 2p(1-q)(pq + (1-p)(1-q)) \\ &\quad - \left(q^2(1-p)^2 + 2q(1-p)(pq + (1-p)(1-q))\right) \\ &= (1 + (1-p)(1-q) + pq) \cdot (p-q)\end{aligned}$$

and so $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(2)}) = (1+(1-p)(1-q)+pq)\cdot(p-q)^2$ (see alternative proof following Proposition 2.4). Thus, the improvement of the case of $k=2$ over the case of $k=1$ is a factor of $(1+(1-p)(1-q)+pq)$, which is greater than 1 unless $\{p,q\}=\{0,1\}$ (where a single sample is as good as k samples, for any $k>1$).

The general case of $k>1$. We now turn to a general analysis of Eq. (17) (and $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$). Specifically, we consider the increase in the value of Eq. (17) when going from k to $k+1$; that is, we define

$$\Delta_{(p,q)}(k) \stackrel{\rm def}{=} \mathrm{E}\left[\mathtt{sign}\left(\sum_{i\in[k+1]} T_i\right)\right] - \mathrm{E}\left[\mathtt{sign}\left(\sum_{i\in[k]} T_i\right)\right] \tag{18}$$

and note that $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k+1)}) = \mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) + (p-q)\cdot\Delta_{(p,q)}(k)$.

Proposition 2.4 (the growth of $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$ as a function of k): *For every $k\geq 1$, it holds that $\Delta_{(p,q)}(k) = (p-q)\cdot\Pr[S_k=0]$, where $S_k \stackrel{\rm def}{=} \sum_{i\in[k]} T_i$.*

It follows that $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k+1)}) = \mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) + (p-q)^2\cdot\Pr[S_k=0]$, and so $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k+1)}) \geq \mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$, where equality holds if and only if $\{p,q\}=\{0,1\}$ (when ignoring the case of $p=q$). Proposition 2.4 can also be used to re-establish $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(2)}) = (1+pq+(1-p)(1-q))\cdot(p-q)^2$, since $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(1)}) = (p-q)^2$ and $\Pr[S_1=0] = pq+(1-p)(1-q)$.

Proof: Starting with Eq. (18), we have

$$\begin{aligned}
\Delta_{(p,q)}(k) &= \mathrm{E}[\mathtt{sign}(S_k+T_{k+1})] - \mathrm{E}[\mathtt{sign}(S_k)] \\
&= \sum_{s\in\{-1,0,1\}} \Pr[S_k=s]\cdot\mathrm{E}[\mathtt{sign}(s+T_{k+1}) - \mathtt{sign}(s)] \\
&= \Pr[S_k=0]\cdot(\Pr[T_{k+1}=1] - \Pr[T_{k+1}=-1]) \\
&\quad +\Pr[S_k=-1]\cdot\Pr[T_{k+1}=1] - \Pr[S_k=1]\cdot\Pr[T_{k+1}=-1]
\end{aligned}$$

By symmetry (e.g., consider the case of $k=1$), it is rather self-evident that $\Pr[S_k=-1]\cdot\Pr[T_{k+1}=1] = \Pr[S_k=1]\cdot\Pr[T_{k+1}=-1]$, yet we provide a detailed proof next.

$$\begin{aligned}
\Pr[S_k=-1]\cdot\Pr[T_{k+1}=1] &= p(1-q)\cdot\sum_{j=1}^{k} B(p,j-1,k)B(q,j,k) \\
&= p(1-q)\cdot\sum_{j=1}^{k}\binom{k}{j-1}p^{j-1}(1-p)^{k-j+1}\binom{k}{j}q^j(1-q)^{k-j} \\
&= \sum_{j=1}^{k}\binom{k}{j-1}p^{j}(1-p)^{k+1-j}\binom{k}{j}q^j(1-q)^{k-j+1} \\
&= (1-p)q\sum_{j=1}^{k}\binom{k}{j-1}p^{j}(1-p)^{k-j}\binom{k}{j}q^{j-1}(1-q)^{k-j+1} \\
&= (1-p)q\cdot\sum_{j=1}^{k} B(p,j,k)B(q,j-1,k) \\
&= \Pr[S_k=1]\cdot\Pr[T_{k+1}=-1]
\end{aligned}$$

Hence, $\Delta_{(p,q)}(k) = \Pr[S_k = 0] \cdot (\Pr[T_{k+1}=1] - \Pr[T_{k+1}=-1])$, and the claim follows (because $\Pr[T_{k+1}=1] - \Pr[T_{k+1}=-1] = p - q$). ■

Proposition 2.4 yields another expression for $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$:

$$\begin{aligned} \mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) &= \mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(1)}) + (p-q) \cdot \sum_{\ell=1}^{k-1} \Delta_{(p,q)}(\ell) & (19) \\ &= (p-q)^2 + (p-q)^2 \cdot \sum_{\ell=1}^{k-1} \Pr[S_\ell=0] & (20) \end{aligned}$$

Note that for $\{p,q\} = \{0,1\}$ this expression (i.e., Eq. (20)) equals 1 (for any $k \geq 1$), whereas for $p = q$ it equals 0. In all other cases (i.e., $0 < (p-q)^2 < 1$) Eq. (20) grows with k. Using $\Pr[S_\ell=0] = \sum_{j=0}^{\ell} B(p,j,\ell)B(q,j,\ell)$, we get

$$\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) = (p-q)^2 + (p-q)^2 \cdot \sum_{\ell=1}^{k-1}\sum_{j=0}^{\ell} \binom{\ell}{j}^2 (pq)^j((1-p)(1-q))^{\ell-j} \qquad (21)$$

In the special case of $p = 0$, Eq. (21) yields

$$\begin{aligned} \mathcal{V}_{(0,q)}(f_{\mathtt{sd}}^{(k)}) &= q^2 + q^2 \cdot \sum_{\ell=1}^{k-1} (1-q)^\ell \\ &= q^2 + q \cdot \left((1-q) - (1-q)^k\right) \end{aligned}$$

which converges to $q = |p-q|$ when $k \to \infty$. Similarly, $\mathcal{V}_{(1,q)}(f_{\mathtt{sd}}^{(k)})$ converges to $1 - q = |p-q|$ (where $p = 1$). Note that in these cases convergence occurs with $k \gg |p-q|^{-1}$. As we shall see next, in the other cases (i.e., $p, q \in (0,1)$), convergence occurs with $k \gg |p-q|^{-2}$.

Proposition 2.5 (approximating $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$ as a function of k): *For any fixed $p, q \in (0,1)$ and sufficiently large k, it holds that $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) = v \cdot |p-q|$, where $v = \Theta(\sqrt{k}) \cdot |p-q|$ if $k \leq (p-q)^{-2}$ and $v = 1 - \exp(-\Theta((p-q)^2 k))$ otherwise.*

Proof: We shall approximate $\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)})$ by using Eq. (16) (rather than Eq. (21)). Recall that by Eq. (16) we have

$$\mathcal{V}_{(p,q)}(f_{\mathtt{sd}}^{(k)}) = (p-q) \cdot \mathrm{E}[\mathtt{sign}(S_k)] \qquad (22)$$

where $S_k = \sum_{i=1}^{k} T_i$ (and $T_i = X_i - Y_i$). Now, for any fixed $p, q \in (0,1)$ and all sufficiently large k, we approximate $\mathrm{E}[\mathtt{sign}(S_k)]$ by $\mathrm{E}[\mathtt{sign}(\widetilde{S}_k)]$, where $\widetilde{S}_k$ is the normal distribution approximation of S_k; that is,

$$\widetilde{S}_k \stackrel{\text{def}}{=} k \cdot (p-q) + \gamma_{p,q} \cdot \sqrt{k} \cdot \mathrm{N}(0,1), \qquad (23)$$

where $\gamma_{p,q} = \frac{1}{2\pi\cdot(p(1-p)+q(1-q))}$ and $\mathrm{N}(0,1)$ denotes the normal distribution (with mean 0 and variance 1). Turning to the analysis of Eq. (22), we replace $\mathrm{E}[\mathtt{sign}(S_k)]$ by $\mathrm{E}[\mathtt{sign}(\widetilde{S}_k)]$, and use

$$\begin{aligned} \mathrm{E}[\mathtt{sign}(\widetilde{S}_k)] &= \Pr[\widetilde{S}_k > 0] - \Pr[\widetilde{S}_k < 0] & (24) \\ &= 2\Pr[\widetilde{S}_k > 0] - 1. & (25) \end{aligned}$$

Now, we analyze $\Pr[\widetilde{S}_k > 0]$ via

$$\Pr[(p-q)k + \gamma_{p,q}\sqrt{k}\mathrm{N}(0,1) > 0] \;=\; \Pr\left[\mathrm{N}(0,1) > -\frac{p-q}{\gamma_{p,q}} \cdot \sqrt{k}\right] \tag{26}$$

Assuming that $p > q$ we get:

- If $k \geq (p-q)^{-2}$ then $r \stackrel{\rm def}{=} (p-q)\sqrt{k} \geq 1$, and it follows that $\Pr[\mathrm{N}(0,1) > -r/\gamma_{p,q}] > 1 - \exp(-\Theta(r^2))$, where the Theta-notation hides a quadratic dependence on ${\gamma_{p,q}}^{-1}$. So Eq. (25) yields more than $1 - \exp(-\Theta(r^2))$, and $\mathcal{V}_{(p,q)}(f_{\tt sd}^{(k)}) > (1 - \exp(-\Theta(r^2))) \cdot (p-q)$.

- If $k \leq (p-q)^{-2}$ then $\epsilon \stackrel{\rm def}{=} (p-q)\sqrt{k} \leq 1$, and it follows that $\Pr[\mathrm{N}(0,1) > -\epsilon/\gamma_{p,q}] = 0.5 + \Theta(\epsilon)$. So Eq. (25) yields $\Theta(\sqrt{k} \cdot (p-q))$, and $\mathcal{V}_{(p,q)}(f_{\tt sd}^{(k)}) = \Theta(\sqrt{k}) \cdot (p-q)^2$.

A similar analysis applies to $p < q$, where we upper bound $\Pr[\widetilde{S}_k > 0] = \Pr[\mathrm{N}(0,1) > \sqrt{k}(q-p)/\gamma_{p,q}]$, and use Eq. (25). ■

3 Conclusion

The obvious way of using statistical information (e.g., a binary guess that is positively correlated with the correct value) is to amplify the confidence level of the information and use it as if it were certainly correct. The current note studies an alternative method of using statistical information and shows that in some settings using unreliable information directly works quite well. This was demonstrated already in Section 1.2, whereas the rest of this note studies the question of how to make the best use of multiple independent copies of such statistical information.

Acknowledgments

We are grateful to Ofer Zeitouni and Dana Ron for useful discussions.

References

[1] O. Goldreich. *Foundation of Cryptography – Basic Tools.* Cambridge University Press, 2001.

[2] O. Goldreich. *Foundation of Cryptography – Basic Applications.* Cambridge University Press, 2004.