AN AXIOMATIC DERIVATION OF THE INFORMATION MEASURE

Let X and Y be discrete random variables with respective alphabets $\mathcal{X}$ and $\mathcal{Y}$. It may help to think of X and Y as representing the input and output of some digital communication system. We are interested in quantifying the amount of information that observation of the occurrence of the event $[Y = y]$ provides about whether or not the event $[X = x]$ also has occurred. We denote this quantity by $I(x, y)$. We assume knowledge of the joint distribution $p(x, y) = Pr[X = x, Y = y]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. This, of course, provides us with knowledge of the associated marginal distributions $\{p(x), x \in \mathcal{X}\}$ and $\{q(y), y \in \mathcal{Y}\}$ and conditional distributions $\{p(x|y)\}$ and $\{q(y|x)\}$.

We now introduce four postulates, or requirements, that most people consider it reasonable that $I(x, y)$ should obey. After each postulate is introduced, we name it and try to describe the motivation underlying it.

Postulate A. There exists a function $F(\alpha, \beta)$ such that $I(x, y) = F(\alpha, \beta)|_{\alpha=p(x),\beta=p(x|y)}$

The idea behind this postulate is that $[Y = y]$ can convey information about $[X = x]$ only by virtue of the fact that it changes the probability of occurrence of $[X = x]$ from its *apriori* value $p(x)$ to its *aposteriori* value $p(x|y)$. We call Postulate A the **Bayesian Postulate** because it imbeds information into the Bayesian framework for reasoning probabilistically from observations back to their possible causes.

Postulate B. The partial derivatives of $F(\alpha, \beta)$ exist.

That is, $F_1(\alpha, \beta) = \frac{\partial}{\partial \alpha} F(\alpha, \beta)$ and $F_2(\alpha, \beta) = \frac{\partial}{\partial \beta} F(\alpha, \beta)$ exist for $0 \leq \alpha, \beta \leq 1$. We call Postulate A the **Smoothness Postulate**. Since differentiability implies continuity, the Smoothness Postulate implies among other things that an infinitesmal perturbation in the prior or posterior probability of occurrence of $[X = x]$ cannot result in a discontinuous jump in our information measure.

Postulate C. $F(\alpha, \gamma) = F(\alpha, \beta) + F(\beta, \gamma)$, $0 \leq \alpha, \beta, \gamma \leq 1$.

The reasoning underlying Postulate C is that, if y were a vector with two components, say $y = (w, z)$, then the information provided by observing its occurrence would have to be the sum of that provided by observing w and that provided by then observing z. In the first of these two steps the information that $[W = w]$ provides about whether or not $[X = x]$ is $F(p(x), p(x|w))$. Once this information has been provided, the original prior probability $p(x)$ of the event $[X = x]$ is replaced by $p(x|w)$. After $[Z = z]$ subsequently is observed, the posterior probability of occurrence of $[X = x]$ then becomes $p(x|w, z)$, so the additional information provided must be $F(p(x|w), p(x|w, z))$. We conclude that $F(p(x), p(x|w, z)) = F(p(x), p(x|w)) + F(p(x|w), p(x|w, z))$. Since $p(x), p(x|w)$ and $p(x|w, z)$ can range over any numbers in the unit cube in various examples, we are led to Postulate C, which we call the **Successive Revelation Postulate.**

Postulate D. $F(\alpha\gamma, \beta\delta) = F(\alpha, \beta) + F(\gamma, \delta)$, $0 \leq \alpha, \beta, \gamma, \delta \leq 1$.

The motivation behind Postulate D is that, if we have two independent experiments, one with input X and output Y and the other with input U and output V, then the information that observation of the combined output event $[Y = y, V = v]$ provides about whether or not the combined input event $[x = x, U = u]$ occurred should be the sum of that which $[Y = y]$ provides about whether or not $[X = x]$ and that which $[V = v]$ provides about whether or not $[U = u]$. Whenever the (X, Y)

and (U, V) experiments are independent, though, the joint prior probability is $p(x, u) = p(x)p(u)$ and the joint posterior probability is $p(x, u|y, v) = p(x|y)p(u|v)$. Hence, we require that

$$F(p(x)p(u), p(x|y)p(u|v)) = F(p(x), p(x|y)) + F(p(u), p(u|v)).$$

Since $p((x), p(x|y), p(u)$ and $p(u|v)$ can assume any values in $[0, 1]^4$ in various examples, we are led to Postulate D, which we call the **Independence Additivity Postulate.**

Deriving the Expression for $I(x, y)$

We now use Postulates A-D to derive Shannon's logarithmic measure of information. [N.B. This is **not** the way Shannon arrived at this way of measuring information. The real justification for his definition of information resides in the source and channel coding theorems that we shall establish later.] For convenience of reference all four postulates are gathered together here.

Postulate A. There exists a function $F(\alpha, \beta)$ such that $I(x, y) = F(\alpha, \beta)|_{\alpha=p(x),\beta=p(x|y)}$

Postulate B. The partial derivatives of $F(\alpha, \beta)$ exist.

Postulate C. $F(\alpha, \gamma) = F(\alpha, \beta) + F(\beta, \gamma)$, $0 \leq \alpha, \beta, \gamma \leq 1$.

Postulate D. $F(\alpha\gamma, \beta\delta) = F(\alpha, \beta) + F(\gamma, \delta)$, $0 \leq \alpha, \beta, \gamma, \delta \leq 1$.

Because of B we may take the partial derivative of both sides of C with respect to β. However, β does not appear on the left hand side of C, so we get

$$0 = F_2(\alpha, \beta) + F_1(\beta, \gamma),$$

or equivalently, $F_2(\alpha, \beta) = -F_1(\beta, \gamma)$. It follows that $F_2(\alpha, \beta)$ cannot vary with α because α does not appear in $F_1(\beta, \gamma)$. That is, $F_2(\alpha, \beta)$ is actually a function only of β which we shall denote by $G'(\beta)$. We have

$$F_2(\alpha, \beta) = -F_1(\beta, \gamma) = G'(\beta).$$

Next observe that if we take the indefinite integral of $F_2(\alpha, \beta)$ with respect to β, we have to get back $F(\alpha, \beta)$ plus a constant of integration, $C = C(\alpha)$, where we have been careful to allow for the fact that the constant may depend on the other argument α in $F(\alpha, \beta)$. That is,

$$\int F_2(\alpha, \beta) d\beta = F(\alpha, \beta) + C(\alpha).$$

[Check this by taking the partial with respect to β and verifying that you get the identity $F_2(\alpha, \beta) = F_2(\alpha, \beta)$.] Hence, we may write

$$\int G'(\beta) d\beta = G(\beta) = F(\alpha, \beta) + C(\alpha),$$

so $F(\alpha, \beta) = G(\beta) - C(\alpha)$. Putting this into C, we get

$$G(\gamma) - C(\alpha) = G(\beta) - C(\alpha) + G(\gamma) - C(\beta),$$

which tells us that $G(\beta) = C(\beta)$. Accordingly,

$$F(\alpha, \beta) = G(\beta) - G(\alpha).$$

Our problem of discovering the functional form of $F(\alpha, \beta)$, a function of two variables, thus has been reduced to that of finding the function $G(\cdot)$ of a single variable.

Now we use Postulate D re-expressed in terms of $G(\cdot)$, namely

$$G(\beta\delta) - G(\alpha\gamma) = G(\beta) - G(\alpha) + G(\delta) - G(\gamma).$$

Taking the derivative of this with respect to δ gives

$$\beta G'(\beta\delta) = G'(\delta).$$

In the limit as $\delta \to 1$ this becomes

$$\beta G'(\beta) = G'(1) = K,$$

where K is a constant. This tells us that

$$G'(\beta) = K/\beta,$$

from which we deduce that

$$G(\beta) = K\ln(\beta) + C,$$

where C is another constant. It follows that

$$F(\alpha, \beta) = K\ln(\beta) + C - K\ln(\alpha) - C,$$

or

$$F(\alpha, \beta) = K\ln(\frac{\beta}{\alpha}).$$

Referring to Postulate A, we conclude that

$$I(x, y) = K\ln(\frac{p(x|y)}{p(x)}).$$

The constant K determines the unit of information. If we set $K = 1$, then I is measured in "nats." It is more common to set $K = \log_2(e) = 1.443$, in which case we say I is measured in "bits" and write

$$I(x, y) = \log_2(\frac{p(x|y)}{p(x)}) \text{ bits.}$$