47. *Principle of the Minimum Entropy in Information Theory*

By Hirosi ITÔ

Department of Applied Physics, Faculty of Engineering, Osaka University

(Comm. by M. MASIMA, M.J.A., May 13, 1953)

1. Introduction

In the previous paper, we have found that the proper representation of the entropy of the continuous information was difficult and that the uncertainty relation had to be taken into account in order to complete this theory. However, the characteristics of the ensemble are given by its autocorrelation function or its spectral density, so that it is desirable to use them to represent its entropy. The key to solve this problem seems to be given in Shannon's paper where "the entropy loss in linear filter" is discussed[1]. This calculations are based on the theory of the filter which is relatively simpler than the theory of Wiener's R.M.S. criterion.

Although Wiener's theory is brilliant and strictly constructed, it is not in vain to rewrite it from the information theory. Because the prediction or filtering is to reduce the uncertainty of the system and hence, there is some hope to translate the idea of the R.M.S. criterion into the information theoretical representation.

2. Entropy of the Ensemble

Shannon has derived the formula representing the entropy loss which occurred when the ensemble passed through a filter with characteristic $k(\omega)$. It is written as

$$H_0 = H_I + \frac{1}{2\pi W}\int_W \log |k(\omega)|^2 \, d\omega \,, \tag{2. 1}$$

where H_I is the entropy of the input per degree of freedom and H_0 is that of the output.

This relation has been derived from the formula

$$f_0(t) = \int_0^\infty f_I(t-\tau)\, dK(\tau) \,, \tag{2. 2}$$

where $f_I(t)$ and $f_0(t)$ are input and output signal respectively. Fourier transform of (2.2) gives

$$A_0(\omega) = A_I(\omega)\, k(\omega) \,. \tag{2. 3}$$

$k(\omega)$ is given by

$$\int_0^\infty e^{-i\omega t} dK(t) = k(\omega) \,.$$

The complex conjugate relation of (2.3) is similarly

$$\overline{A_0(\omega)} = \overline{A_I(\omega)}\ \overline{k(\omega)}\,. \tag{2.4}$$

Combining two relations (2.3) and (2.4), we get

$$|A_0(\omega)|^2 = |A_I(\omega)|^2 |k(\omega)|^2$$

or

$$\Phi_0(\omega) = \Phi_I(\omega)\, |k(\omega)|^2\,. \tag{2.5}$$

Substituting (2.5) into (2.1), we may put

$$H_0 - H_I = \frac{1}{2\pi W}\int_W (\log \Phi_0(\omega) - \log \Phi_I(\omega))\, d\omega\,.$$

From this relation we may generally conclude that next relation will exist:

$$H = \frac{1}{2\pi W}\int_W \log \Phi(\omega)\, d\omega + C\,, \tag{2.6}$$

where C is an arbitrary constant. This result cannot be said to represent the entropy of the ensemble exactly. Because, there is no information about the phase in this formula. However, the exact representation is very difficult and this problem will be left in the future.

3. Wiener's R.M.S. Criterion and Principle of the Minimum Entropy

The R.M.S.[2] method which was developed by N. Wiener is very powerful for the prediction or filtering of the time series. The principle of his method is as follows: We must choose a proper operator $K(t)$ so as to minimize the average power of the error signal

$$\varepsilon(t) = f(t+\alpha) - \int_0^\infty \{f(t-\tau) + g(t-\tau)\}\, dK(\tau)\,, \tag{3.1}$$

where α is the delay time. That is:

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \varepsilon(t)^2 dt = \text{minimum}.$$

We can translate this method into the information theoretical one; if the operator $K(t)$ is chosen so that the entropy of the error signal may be minimum, we can make an optimum prediction or filtering. While this principle seems to be very interesting and reasonable, we cannot obtain the Wiener's result because of the incompleteness of the representation of the entropy. However, it is easily seen for this principle to be almost exact even if using the expression (2.6). Let us put the apriori entropy H_1 and a posteriori one H_2. The information which is obtained after the observation is[3]

$$I = H_1 - H_2.$$

By (2.6), it becomes

$$I=\frac{1}{2\pi W}\int_W \log\left(\frac{\Phi_1(\omega)}{\Phi_2(\omega)}\right)d\omega\,.$$

If we put $$\Phi_1=\Phi_2+\delta\Phi_2\,,$$
the corresponding information δI is

$$\frac{1}{2\pi W}\int_W \log\left(1+\frac{\delta\Phi_2(\omega)}{\Phi_2(\omega)}\right)d\omega\,. \tag{3.2}$$

When $\delta\Phi_2(\omega)=0$, we have the maximum information from this ensemble. Now we consider the following experiment: When some operater $K_1(t)$ is chosen, at first. The corresponding error function, which will still have some entropy to transmit some information, will be obtained. Next, let this obtained error signal pass through the filter, where we shall obtain more information. Repeating these experiments, we shall have no more information at last. There may still exist some entropy, but we cannot reduce it any more. Therefore we may minimize the error spectrum by (3.2) where $\Phi_2(\omega)$ is replaced by $\Phi_\varepsilon(\omega)$. Now, in order to apply this principle to the prediction or filtering, it must be necessary to choose the operator $K(t)$ which minimizes the erorr spectral density by means of (3.1). Now the autocorrelation function of the error signal is

$$\begin{aligned}
&\overline{\varepsilon(t+s)\,\overline{\varepsilon(t)}}\\
&=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\Big\{f(t+\alpha+s)-\int_0^\infty[f(t-\tau+s)+g(t-\tau+s)]\,dK(\tau)\Big\}\\
&\qquad\qquad\times\Big\{\overline{f(t+\alpha)}-\int_0^\infty[\overline{f(t-\tau)}+\overline{g(t-\tau)}]\,d\overline{K(\tau)}\Big\}\\
&=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\Big\{f(t+\alpha+s)\overline{f(t+\alpha)}-f(t+\alpha+s)\int_0^\infty[\overline{f(t-\sigma)}+\overline{g(t-\sigma)}]\,d\overline{K(\sigma)}\\
&\qquad-\overline{f(t+\alpha)}\int_0^\infty[f(t-\tau+s)+g(t-\tau+s)]\,dK(\tau)\\
&\qquad+\int_0^\infty dK(\tau)\int_0^\infty d\overline{K(\sigma)}[f(t-\tau+s)+g(t-\tau+s)]\times[\overline{f(t-\sigma)}+\overline{g(t-\sigma)}]\Big\}\,dt\\
&=\varphi_{11}(s)-\int_0^\infty\{\varphi_{11}(\alpha+s+\sigma)+\varphi_{12}(\alpha+s+\sigma)\}\,d\overline{K(\sigma)}\\
&\qquad-\int_0^\infty\{\overline{\varphi_{11}(\alpha+\sigma-s)}+\overline{\varphi_{12}(\alpha+\sigma-s)}\}\,dK(\sigma)\\
&\qquad+\int_0^\infty dK(\sigma)\int_0^\infty d\overline{K(\tau)}\,\varphi(\tau-\sigma+s),
\end{aligned}$$

where
$$\varphi_{11}(\tau)=\frac{1}{2T}\int_{-T}^{T}f(t+\tau)\overline{f(t)}\,dt$$
$$\varphi_{12}(\tau)=\frac{1}{2T}\int_{-T}^{T}f(t+\tau)\,\overline{g(t)}\,dt,$$
and
$$\varphi(\tau)=\varphi_{11}(\tau)+\varphi_{12}(\tau)+\overline{\varphi_{1}(-\tau)}+\varphi_{22}(\tau). \tag{3.3}$$

The spectral density of the error signal is

$$
\begin{aligned}
(\Phi_\varepsilon\omega) &= \int_{-\infty}^{\infty} \overline{\varepsilon(t+\tau)\,\overline{\varepsilon(t)}}\, e^{-i\omega\tau}\, d\tau \\
&= \Phi_{11}(\omega) - \int_0^\infty {}^{i\omega\sigma} d\overline{K(\sigma)} \int_{-\infty}^{\infty} \{\varphi_{11}(\alpha+s+\sigma)+\varphi_{12}(\alpha+s+\sigma)\}\, e^{-i\omega(s+\sigma+\alpha)} \\
&\qquad \times e^{i\omega\alpha} ds - \int_0^\infty e^{-i\omega\sigma} dK(\sigma) \int_{-\infty}^{\infty} \{\varphi_{11}(\overline{\alpha+\sigma-s}) + \varphi_{12}(\overline{\alpha+\sigma-s})\} \\
&\qquad \times e^{i\omega(\alpha+\sigma-s)}\, e^{-i\omega\alpha}\, ds \\
&\qquad + \int_0^\infty e^{-i\omega\sigma} dK(\sigma) \int_0^\infty e^{i\omega\tau} d\overline{K(\tau)} \int_{-\infty}^{\infty} \varphi(\tau-\sigma+s)\, e^{-i\omega(s+\tau-\sigma)} ds \\
&= \Phi_{11}(\omega) - \overline{k(\omega)}\,(\Phi_{11}(\omega)+\Phi_{12}(\omega))\, e^{i\omega\alpha} - k(\omega)\,(\overline{\Phi_{11}(\omega)} + \overline{\Phi_{12}(\omega)})\, e^{-i\omega\alpha} \\
&\qquad + k(\omega)\overline{k(\omega)}\,\Phi(\omega).
\end{aligned}
$$

By the condition $\overline{\varphi(-\tau)}=\varphi(\tau)$, $\Phi(\omega)$ is real. Therefore we have the same result if we vary $k(\omega)$ or $\overline{k(\omega)}$ independently. Then,

$$\delta\Phi_\varepsilon(\omega) = 0 = k(\omega)\,\Phi(\omega) - (\Phi_{11}(\omega)+\Phi_{12}(\omega))\, e^{-i\alpha\omega}$$

Hence it gives

$$e^{-i\alpha\omega} k(\omega) = \frac{\Phi_{11}(\omega) + \Phi_{12}(\omega)}{\Phi(\omega)}.$$

This is the same result as that obtained by Wiener when α becomes $-\infty$ in the limiting case.

4. Conclusions

The reason why we get the limiting case of Wiener's is that the entropy representation (2.6) is incomplete, because it has no information about the part of the phase. Therefore, although this result is not sufficient, it can be said that this principle will not be incorrect. The extreme value of $\Phi_\varepsilon(\omega)$ is

$$\Phi_{11}(\omega) - |k(\omega)|^2\, \Phi(\omega)$$

and is also the same result as Wiener's. That is, this principle satisfies the condition of the minimum average power of the error signal.

References

1) The Mathematical Theory of Communication, 60.

2) The Extraporation, Interporation, and Smoothing of Stationary Time Series, NDRC Report, Cambridge, Mass., 1942.

3) H. Itô: Proc. Japan Acad., **28**, 187 (1952).