3.17 ESSENTIAL AVERAGE MUTUAL INFORMATION

Yaser S. Abu-Mostafa

California Institute of Technology
Pasadena, CA 91125

Consider two dependent random variables (S, C) and suppose that $\hat{\chi}$ is the optimal estimate of C when only S is known. $I(S; C)$ is a measure of how much S tells us about C, and $I(\hat{\chi}; C)$ is a measure of how much our optimal estimate $\hat{\chi}$ tells us about C. What can we say about $I(\hat{\chi}; C)$ if we know that $I(S; C) = 3$ bits, for example? The optimality of $\hat{\chi}$ suggests that $I(\hat{\chi}; C)$ should also be close to 3 bits. This is what we address in this problem. Let (S, C) be jointly distributed $\sim p(s,c)$, where $S = \{0, \ldots, N-1\}$ and $C = \{0, \ldots, M-1\}$. Let $\hat{\chi} : \{0, \ldots, N-1\} \to \{0, \ldots, M-1\}$ denote an arbitrary function of the outcomes of S. The problem is to estimate the numbers $\alpha(N, M)$ defined by

$$\alpha(N, M) = \inf_{p: I(S;C) > 0} \max_{\hat{\chi} = \hat{C}(S)} \left[\frac{I(\hat{\chi}; C)}{I(S; C)} \right]$$

Since $I(\hat{\chi}; C) \le I(S; C)$ (data processing inequality), $\alpha(N, M) \le 1$. In fact, $\alpha(N, M) < 1$ for N, M as shown in the following example for $\alpha(3, 2)$.

p(s,c)

C \ S	0	1	2
0	1/3	1/6	0
1	0	1/6	1/3

Either $\hat{\chi}(0) = \hat{\chi}(1)$, $\hat{\chi}(1) = \hat{\chi}(2)$, or $\hat{\chi}(2) = \hat{\chi}(0)$ will make $I(\hat{\chi}; C) < I(S; C)$.

Generalizations.

1. We can think of $\hat{\chi}$ in general as a compression of S. This generalizes $\alpha(N, M)$ to $\alpha(N, M, K)$, where $S = \{ 0, \ldots, N-1 \}$, $C = \{ 0, \ldots, M-1 \}$, and $\hat{\chi} : \{ 0, \ldots, N-1 \} \to \{ 0, \ldots, k-1 \}$.

2. To avoid the cases of very weak dependence between S and C, the minimization domain $(I(S; C) > 0)$ can be restricted to $I(S; C) \geq \delta$ or $I(S; C) \geq \varepsilon H(C)$.