Chapter 10

THE BLAHUT-ARIMOTO ALGORITHMS

For a discrete memoryless channel $p(y|x)$, the capacity

$$C = \max_{r(x)} I(X;Y), \tag{10.1}$$

where X and Y are respectively the input and the output of the generic channel and $r(x)$ is the input distribution, characterizes the maximum asymptotically achievable rate at which information can be transmitted through the channel reliably. The expression for C in (10.1) is called a *single-letter characterization* because it depends only the transition matrix of the generic channel but not on the block length n of a code for the channel. When both the input alphabet $\mathcal{X}$ and the output alphabet $\mathcal{Y}$ are finite, the computation of C becomes a finite-dimensional maximization problem.

For an i.i.d. information source $\{X_k, k \geq 1\}$ with generic random variable X, the rate distortion function

$$R(D) = \min_{Q(\hat{x}|x): Ed(X,\hat{X}) \leq D} I(X;\hat{X}) \tag{10.2}$$

characterizes the minimum asymptotically achievable rate of a rate distortion code which reproduces the information source with an average distortion no more than D with respect to a single-letter distortion measure d. Again, the expression for $R(D)$ in (10.2) is a single-letter characterization because it depends only on the generic random variable X but not on the block length n of a rate distortion code. When both the source alphabet $\mathcal{X}$ and the reproduction alphabet $\hat{\mathcal{X}}$ are finite, the computation of $R(D)$ becomes a finite-dimensional minimization problem.

Unless for very special cases, it is not possible to obtain an expression for C or $R(D)$ in closed form, and we have to resort to numerical computation.

However, computing these quantities is not straightforward because the associated optimization problem is nonlinear. In this chapter, we discuss the *Blahut-Arimoto algorithms* (henceforth the BA algorithms), which is an iterative algorithm devised for this purpose.

In order to better understand how and why the BA algorithm works, we will first describe the algorithm in a general setting in the next section. Specializations of the algorithm for the computation of C and $R(D)$ will be discussed in Section 10.2, and convergence of the algorithm will be proved in Section 10.3.

10.1 ALTERNATING OPTIMIZATION

In this section, we describe an alternating optimization algorithm. This algorithm will be specialized in the next section for computing the channel capacity and the rate distortion function.

Consider the double supremum

$$\sup_{\mathbf{u}_1 \in A_1} \sup_{\mathbf{u}_2 \in A_2} f(\mathbf{u}_1, \mathbf{u}_2), \tag{10.3}$$

where A_i is a convex subset of $\Re^{n_i}$ for $i = 1, 2$, and f is a function defined on $A_1 \times A_2$. The function f is bounded from above, and is continuous and has continuous partial derivatives on $A_1 \times A_2$. Further assume that for all $\mathbf{u}_2 \in A_2$, there exists a unique $c_1(\mathbf{u}_2) \in A_1$ such that

$$f(c_1(\mathbf{u}_2), \mathbf{u}_2) = \max_{\mathbf{u}'_1 \in A_1} f(\mathbf{u}'_1, \mathbf{u}_2), \tag{10.4}$$

and for all $\mathbf{u}_1 \in A_1$, there exists a unique $c_2(\mathbf{u}_1) \in A_2$ such that

$$f(\mathbf{u}_1, c_2(\mathbf{u}_1)) = \max_{\mathbf{u}'_2 \in A_2} f(\mathbf{u}_1, \mathbf{u}'_2). \tag{10.5}$$

Let $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ and $A = A_1 \times A_2$. Then (10.3) can be written as

$$\sup_{\mathbf{u} \in A} f(\mathbf{u}). \tag{10.6}$$

In other words, the supremum of f is taken over a subset of $\Re^{n_1+n_2}$ which is equal to the Cartesian product of two convex subsets of $\Re^{n_1}$ and $\Re^{n_2}$, respectively.

We now describe an alternating optimization algorithm for computing f^*, the value of the double supremum in (10.3). Let $\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)})$ for $k \geq 0$ which are defined as follows. Let $\mathbf{u}_1^{(0)}$ be an arbitrarily chosen vector in A_1, and let $\mathbf{u}_2^{(0)} = c_2(\mathbf{u}_1^{(0)})$. For $k \geq 1$, $\mathbf{u}^{(k)}$ is defined by

$$\mathbf{u}_1^{(k)} = c_1(\mathbf{u}_2^{(k-1)}) \tag{10.7}$$

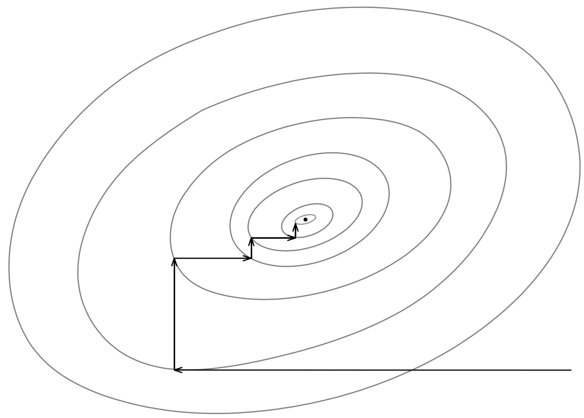

Figure 10.1. Alternating optimization.

and

$$\mathbf{u}_2^{(k)} = c_2(\mathbf{u}_1^{(k)}). \tag{10.8}$$

In other words, $\mathbf{u}_1^{(k)}$ and $\mathbf{u}_2^{(k)}$ are generated in the order $\mathbf{u}_1^{(0)}$, $\mathbf{u}_2^{(0)}$, $\mathbf{u}_1^{(1)}$, $\mathbf{u}_2^{(1)}$, $\mathbf{u}_1^{(2)}$, $\mathbf{u}_2^{(2)}$, $\cdots$, where each vector in the sequence is a function of the previous vector except that $\mathbf{u}_1^{(0)}$ is arbitrarily chosen in A_1. Let

$$f^{(k)} = f(\mathbf{u}^{(k)}). \tag{10.9}$$

Then from (10.4) and (10.5),

$$\begin{aligned} f^{(k)} &= f(\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)}) && (10.10)\\ &\geq f(\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k-1)}) && (10.11)\\ &\geq f(\mathbf{u}_1^{(k-1)}, \mathbf{u}_2^{(k-1)}) && (10.12)\\ &= f^{(k-1)} && (10.13) \end{aligned}$$

for $k \geq 1$. Since the sequence $f^{(k)}$ is non-decreasing, it must converge because f is bounded from above. We will show in Section 10.3 that $f^{(k)} \to f^*$ if f is concave. Figure 10.1 is an illustration of the alternating maximization algorithm, where in this case both n_1 and n_2 are equal to 1, and $f^{(k)} \to f^*$.

The alternating optimization algorithm can be explained by the following analogy. Suppose a hiker wants to reach the summit of a mountain. Starting from a certain point in the mountain, the hiker moves north-south and east-west alternately. (In our problem, the north-south and east-west directions can be multi-dimensional.) In each move, the hiker moves to the highest possible point. The question is whether the hiker can eventually approach the summit starting from any point in the mountain.

Replacing f by $-f$ in (10.3), the double supremum becomes the double infimum

$$\inf_{\mathbf{u}_1 \in A_1} \inf_{\mathbf{u}_2 \in A_2} f(\mathbf{u}_1, \mathbf{u}_2). \tag{10.14}$$

All the previous assumptions on A_1, A_2, and f remain valid except that f is now assumed to be bounded from below instead of bounded from above. The double infimum in (10.14) can be computed by the same alternating optimization algorithm. Note that with f replaced by $-f$, the maximums in (10.4) and (10.5) become minimums, and the inequalities in (10.11) and (10.12) are reversed.

10.2 THE ALGORITHMS

In this section, we specialize the alternating optimization algorithm described in the last section to compute the channel capacity and the rate distortion function. The corresponding algorithms are known as the BA algorithms.

10.2.1 CHANNEL CAPACITY

We will use $\mathbf{r}$ to denote an input distribution $r(x)$, and we write $\mathbf{r} > 0$ if $\mathbf{r}$ is strictly positive, i.e., $r(x) > 0$ for all $x \in \mathcal{X}$. If $\mathbf{r}$ is not strictly positive, we write $\mathbf{r} \geq 0$. Similar notations will be introduced as appropriate.

LEMMA 10.1 *Let $r(x)p(y|x)$ be a given joint distribution on $\mathcal{X} \times \mathcal{Y}$ such that $\mathbf{r} > 0$, and let $\mathbf{q}$ be a transition matrix from $\mathcal{Y}$ to $\mathcal{X}$. Then*

$$\max_{\mathbf{q}} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} = \sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{r(x)}, \tag{10.15}$$

where the maximization is taken over all $\mathbf{q}$ such that

$$q(x|y) = 0 \quad \textit{if and only if} \quad p(y|x) = 0, \tag{10.16}$$

and

$$q^*(x|y) = \frac{r(x)p(y|x)}{\sum_{x'} r(x')p(y|x')}, \tag{10.17}$$

i.e., the maximizing $\mathbf{q}$ is the which corresponds to the input distribution $\mathbf{r}$ and the transition matrix $p(y|x)$.

In (10.15) and the sequel, we adopt the convention that the summation is taken over all x and y such that $r(x) > 0$ and $p(y|x) > 0$. Note that the right hand side of (10.15) gives the mutual information $I(X;Y)$ when $\mathbf{r}$ is the input distribution for the generic channel $p(y|x)$.

Proof Let

$$w(y) = \sum_{x'} r(x')p(y|x') \tag{10.18}$$

in (10.17). We assume with loss of generality that for all $y \in \mathcal{Y}$, $p(y|x) > 0$ for some $x \in \mathcal{X}$. Since $\mathbf{r} > 0$, $w(y) > 0$ for all y, and hence $q^*(x|y)$ is well-defined. Rearranging (10.17), we have

$$r(x)p(y|x) = w(y)q^*(x|y). \tag{10.19}$$

Consider

$$\begin{aligned}
&\sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{r(x)} - \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} \\
&= \sum_x \sum_y r(x)p(y|x) \log \frac{q^*(x|y)}{q(x|y)} &(10.20)\\
&= \sum_y \sum_x w(y)q^*(x|y) \log \frac{q^*(x|y)}{q(x|y)} &(10.21)\\
&= \sum_y w(y) \sum_x q^*(x|y) \log \frac{q^*(x|y)}{q(x|y)} &(10.22)\\
&= \sum_y w(y) D(q^*(x|y) \| q(x|y)) &(10.23)\\
&\geq 0, &(10.24)
\end{aligned}$$

where (10.21) follows from (10.19), and the last step is an application of the divergence inequality. Then the proof is completed by noting in (10.17) that $\mathbf{q}^*$ satisfies (10.16) because $\mathbf{r} > 0$. □

THEOREM 10.2 *For a discrete memoryless channel $p(y|x)$,*

$$C = \sup_{\mathbf{r}>0} \max_{\mathbf{q}} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)}, \tag{10.25}$$

where the maximization is taken over all $\mathbf{q}$ which satisfies (10.16).

Proof Let $I(\mathbf{r}, \mathbf{p})$ denote the mutual information $I(X;Y)$ when $\mathbf{r}$ is the input distribution for the generic channel $p(y|x)$. Then we can write

$$C = \max_{\mathbf{r}\geq 0} I(\mathbf{r}, \mathbf{p}). \tag{10.26}$$

Let $\mathbf{r}^*$ achieves C. If $\mathbf{r}^* > 0$, then

$$\begin{aligned}
C &= \max_{\mathbf{r}\geq 0} I(\mathbf{r}, \mathbf{p}) &(10.27)\\
&= \max_{\mathbf{r}>0} I(\mathbf{r}, \mathbf{p}) &(10.28)\\
&= \max_{\mathbf{r}>0} \max_{\mathbf{q}} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} &(10.29)\\
&= \sup_{\mathbf{r}>0} \max_{\mathbf{q}} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)}, &(10.30)
\end{aligned}$$

where (10.29) follows from Lemma 10.1 (and the maximization is over all $\mathbf{q}$ which satisfies (10.16)).

Next, we consider the case when $\mathbf{r}^* \geq 0$. Since $I(\mathbf{r}, \mathbf{p})$ is continuous in $\mathbf{r}$, for any $\epsilon > 0$, there exists $\delta > 0$ such that if

$$\|\mathbf{r} - \mathbf{r}^*\| < \delta, \tag{10.31}$$

then

$$C - I(\mathbf{r}, \mathbf{p}) < \epsilon, \tag{10.32}$$

where $\|\mathbf{r} - \mathbf{r}^*\|$ denotes the Euclidean distance between $\mathbf{r}$ and $\mathbf{r}^*$. In particular, there exists $\tilde{\mathbf{r}} > 0$ which satisfies (10.31) and (10.32). Then

$$\begin{aligned} C &= \max_{\mathbf{r} \geq 0} I(\mathbf{r}, \mathbf{p}) && (10.33)\\ &\geq \sup_{\mathbf{r} > 0} I(\mathbf{r}, \mathbf{p}) && (10.34)\\ &\geq I(\tilde{\mathbf{r}}, \mathbf{p}) && (10.35)\\ &> C - \epsilon, && (10.36) \end{aligned}$$

where the last step follows because $\tilde{\mathbf{r}}$ satisfies (10.32). Thus we have

$$C - \epsilon < \sup_{\mathbf{r} > 0} I(\mathbf{r}, \mathbf{p}) \leq C. \tag{10.37}$$

Finally, by letting $\epsilon \to 0$, we conclude that

$$C = \sup_{\mathbf{r} > 0} I(\mathbf{r}, \mathbf{p}) = \sup_{\mathbf{r} > 0} \max_{\mathbf{q}} \sum_x \sum_y r(x) p(y|x) \log \frac{q(x|y)}{r(x)}. \tag{10.38}$$

This accomplishes the proof. □

Now for the double supremum in (10.3), let

$$f(\mathbf{r}, \mathbf{q}) = \sum_x \sum_y r(x) p(y|x) \log \frac{q(x|y)}{r(x)}, \tag{10.39}$$

with $\mathbf{r}$ and $\mathbf{q}$ playing the roles of $\mathbf{u}_1$ and $\mathbf{u}_2$, respectively. Let

$$A_1 = \left\{ (r(x), x \in \mathcal{X}) : \ r(x) > 0 \text{ and } \textstyle\sum_x r(x) = 1 \right\}, \tag{10.40}$$

and

$$\begin{aligned} A_2 = \ & \{ (q(x|y), (x, y) \in \mathcal{X} \times \mathcal{Y}) : \ q(x|y) > 0 \\ & \text{if } p(x|y) > 0, \ q(x|y) = 0 \text{ if } p(y|x) = 0, \\ & \text{and } \textstyle\sum_x q(x|y) = 1 \text{ for all } y \in \mathcal{Y} \}. \end{aligned} \tag{10.41}$$

Then A_1 is a subset of $\Re^{|\mathcal{X}|}$ and A_2 is a subset of $\Re^{|\mathcal{X}||\mathcal{Y}|}$, and it is readily checked that both A_1 and A_2 are convex. For all $\mathbf{r} \in A_1$ and $\mathbf{q} \in A_2$, by Lemma 10.1,

$$\begin{aligned} f(\mathbf{r}, \mathbf{q}) &= \sum_x \sum_y r(x) p(y|x) \log \frac{q(x|y)}{r(x)} && (10.42) \\ &\leq \sum_x \sum_y r(x) p(y|x) \log \frac{q^*(x|y)}{r(x)} && (10.43) \\ &= I(X;Y) && (10.44) \\ &\leq H(X) && (10.45) \\ &\leq \log |\mathcal{X}|. && (10.46) \end{aligned}$$

Thus f is bounded from above. Since for all $\mathbf{q} \in A_2$, $q(x|y) = 0$ for all x and y such that $p(x|y) = 0$, these components of $\mathbf{q}$ are degenerated. In fact, these components of $\mathbf{q}$ do not appear in the definition of $f(\mathbf{r}, \mathbf{q})$ in (10.39), which can be seen as follows. Recall the convention that the double summation in (10.39) is over all x and y such that $r(x) > 0$ and $p(y|x) > 0$. If $q(x|y) = 0$, then $p(y|x) = 0$, and hence the corresponding term is not included in the double summation. Therefore, it is readily seen that f is continuous and has continuous partial derivatives on A because all the probabilities involved in the double summation in (10.39) are strictly positive. Moreover, for any given $\mathbf{r} \in A_1$, by Lemma 10.1, there exists a unique $\mathbf{q} \in A_2$ which maximizes f. It will be shown shortly that for any given $\mathbf{q} \in A_2$, there also exists a unique $\mathbf{r} \in A_1$ which maximizes f.

The double supremum in (10.3) now becomes

$$\sup_{\mathbf{r} \in A_1} \sup_{\mathbf{q} \in A_2} \sum_x \sum_y r(x) p(y|x) \log \frac{q(x|y)}{r(x)}, \tag{10.47}$$

which by Theorem 10.2 is equal to C, where the supremum over all $\mathbf{q} \in A_2$ is in fact a maximum. We then apply the alternating optimization algorithm in the last section to compute C. First, we arbitrarily choose a *strictly positive* input distribution in A_1 and let it be $\mathbf{r}^{(0)}$. Then we define $\mathbf{q}^{(0)}$ and in general $\mathbf{q}^{(k)}$ for $k \geq 0$ by

$$q^{(k)}(x|y) = \frac{r^{(k)}(x) p(y|x)}{\sum_{x'} r^{(k)}(x') p(y|x')} \tag{10.48}$$

in view of Lemma 10.1. In order to define $\mathbf{r}^{(1)}$ and in general $\mathbf{r}^{(k)}$ for $k \geq 1$, we need to find the $\mathbf{r} \in A_1$ which maximizes f for a given $\mathbf{q} \in A_2$, where the constraints on $\mathbf{r}$ are

$$\sum_x r(x) = 1 \tag{10.49}$$

and

$$r(x) > 0 \quad \text{for all } x \in \mathcal{X}. \tag{10.50}$$

We now use the method of Lagrange multipliers to find the best $\mathbf{r}$ by ignoring temporarily the positivity constraints in (10.50). Let

$$J = \sum_x \sum_y r(x) p(y|x) \log \frac{q(x|y)}{r(x)} - \lambda \sum_x r(x). \tag{10.51}$$

For convenience sake, we assume that the logarithm is the natural logarithm. Differentiating with respect to $r(x)$ gives

$$\frac{\partial J}{\partial r(x)} = \sum_y p(y|x) \log q(x|y) - \log r(x) - 1 - \lambda. \tag{10.52}$$

Upon setting $\frac{\partial J}{\partial r(x)} = 0$, we have

$$\log r(x) = \sum_y p(y|x) \log q(x|y) - 1 - \lambda, \tag{10.53}$$

or

$$r(x) = e^{-(\lambda+1)} \prod_y q(x|y)^{p(y|x)}. \tag{10.54}$$

By considering the normalization constraint in (10.49), we can eliminate λ and obtain

$$r(x) = \frac{\prod_y q(x|y)^{p(y|x)}}{\sum_{x'} \prod_y q(x'|y)^{p(y|x')}}. \tag{10.55}$$

The above product is over all y such that $p(y|x) > 0$, and $q(x|y) > 0$ for all such y. This implies that both the numerator and the denominator on the right hand side above are positive, and therefore $r(x) > 0$. In other words, the $\mathbf{r}$ thus obtained happen to satisfy the positivity constraints in (10.50) although these constraints were ignored when we set up the Lagrange multipliers. We will show in Section 10.3.2 that f is concave. Then $\mathbf{r}$ as given in (10.55), which is unique, indeed achieves the maximum of f for a given $\mathbf{q} \in A_2$ because $\mathbf{r}$ is in the interior of A_1. In view of (10.55), we define $\mathbf{r}^{(k)}$ for $k \geq 1$ by

$$r^{(k)}(x) = \frac{\prod_y q^{(k-1)}(x|y)^{p(y|x)}}{\sum_{x'} \prod_y q^{(k-1)}(x'|y)^{p(y|x')}}. \tag{10.56}$$

The vectors $\mathbf{r}^{(k)}$ and $\mathbf{q}^{(k)}$ are defined in the order $\mathbf{r}^{(0)}$, $\mathbf{q}^{(0)}$, $\mathbf{r}^{(1)}$, $\mathbf{q}^{(1)}$, $\mathbf{r}^{(2)}$, $\mathbf{q}^{(2)}, \cdots$, where each vector in the sequence is a function of the previous vector except that $\mathbf{r}^{(0)}$ is arbitrarily chosen in A_1. It remains to show by induction that $\mathbf{r}^{(k)} \in A_1$ for $k \geq 1$ and $\mathbf{q}^{(k)} \in A_2$ for $k \geq 0$. If $\mathbf{r}^{(k)} \in A_1$, i.e., $\mathbf{r}^{(k)} > 0$,

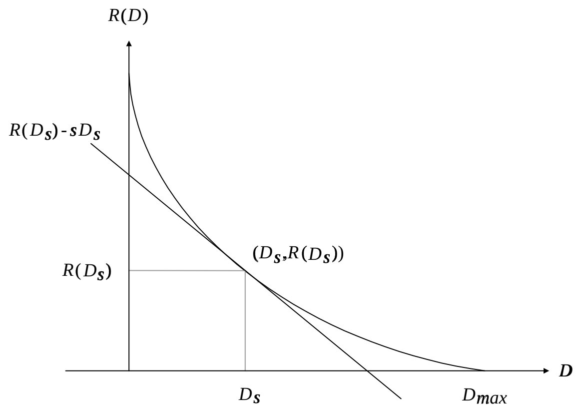

Figure 10.2. A tangent to the $R(D)$ curve with slope equal to s.

then we see from (10.48) that $q^{(k)}(x|y) = 0$ if and only if $p(x|y) = 0$, i.e., $\mathbf{q}^{(k)} \in A_2$. On the other hand, if $\mathbf{q}^{(k)} \in A_2$, then we see from (10.56) that $\mathbf{r}^{(k+1)} > 0$, i.e., $\mathbf{r}^{(k+1)} \in A_2$. Therefore, $\mathbf{r}^{(k)} \in A_1$ and $\mathbf{q}^{(k)} \in A_2$ for all $k \geq 0$. Upon determining $(\mathbf{r}^{(k)}, \mathbf{q}^{(k)})$, we can compute $f^{(k)} = f(\mathbf{r}^{(k)}, \mathbf{q}^{(k)})$ for all k. It will be shown in Section 10.3 that $f^{(k)} \to C$.

10.2.2 THE RATE DISTORTION FUNCTION

This discussion in this section is analogous to the discussion in Section 10.2.1. Some of the details will be omitted for brevity.

For all problems of interest, $R(0) > 0$. Otherwise, $R(D) = 0$ for all $D \geq 0$ since $R(D)$ is nonnegative and non-increasing. Therefore, we assume without loss of generality that $R(0) > 0$.

We have shown in Corollary 9.19 that if $R(0) > 0$, then $R(D)$ is strictly decreasing for $0 \leq D \leq D_{max}$. Since $R(D)$ is convex, for any $s \leq 0$, there exists a point on the $R(D)$ curve for $0 \leq D \leq D_{max}$ such that the slope of a tangent[1] to the $R(D)$ curve at that point is equal to s. Denote such a point on the $R(D)$ curve by $(D_s, R(D_s))$, which is not necessarily unique. Then this tangent intersects with the ordinate at $R(D_s) - sD_s$. This is illustrated in Figure 10.2.

Let $I(\mathbf{p}, \mathbf{Q})$ denote the mutual information $I(X, \hat{X})$ and $D(\mathbf{p}, \mathbf{Q})$ denote the expected distortion $Ed(X, \hat{X})$ when $\mathbf{p}$ is the distribution for X and $\mathbf{Q}$ is the transition matrix from $\mathcal{X}$ to $\hat{\mathcal{X}}$ defining $\hat{X}$. Then for any $\mathbf{Q}$, $(I(\mathbf{p}, \mathbf{Q}), D(\mathbf{p}, \mathbf{Q}))$ is a point in the rate distortion region, and the line with slope s passing through

[1]We say that a line is a tangent to the $R(D)$ curve if it touches the $R(D)$ curve from below.

$(I(\mathbf{p},\mathbf{Q}), D(\mathbf{p},\mathbf{Q}))$ intersects the ordinate at $I(\mathbf{p},\mathbf{Q}) - sD(\mathbf{p},\mathbf{Q})$. Since the $R(D)$ curve defines the boundary of the rate distortion region and it is above the tangent in Figure 10.2, we see that

$$R(D_s) - sD_s = \min_{\mathbf{Q}} [I(\mathbf{p},\mathbf{Q}) - sD(\mathbf{p},\mathbf{Q})]. \tag{10.57}$$

For each $s \leq 0$, if we can find a $\mathbf{Q}_s$ which achieves the above minimum, then the line passing through $(0, I(\mathbf{p},\mathbf{Q}_s) - sD(\mathbf{p},\mathbf{Q}_s))$, i.e., the tangent in Figure 10.2, gives a tight lower bound on the $R(D)$ curve. In particular, if $(R(D_s), D_s)$ is unique,

$$D_s = D(\mathbf{p},\mathbf{Q}_s) \tag{10.58}$$

and

$$R(D_s) = I(\mathbf{p},\mathbf{Q}_s). \tag{10.59}$$

By varying over all $s \leq 0$, we can then trace out the whole $R(D)$ curve. In the rest of the section, we will devise an iterative algorithm for the minimization problem in (10.57).

LEMMA 10.3 *Let $p(x)Q(\hat{x}|x)$ be a given joint distribution on $\mathcal{X} \times \hat{\mathcal{X}}$ such that $\mathbf{Q} > 0$, and let $\mathbf{t}$ be any distribution on $\hat{\mathcal{X}}$ such that $\mathbf{t} > 0$. Then*

$$\min_{\mathbf{t}>0} \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} = \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t^*(\hat{x})}, \tag{10.60}$$

where

$$t^*(\hat{x}) = \sum_x p(x)Q(\hat{x}|x), \tag{10.61}$$

i.e., the minimizing $t(\hat{x})$ is the distribution on $\hat{\mathcal{X}}$ corresponding to the input distribution $\mathbf{p}$ and the transition matrix $\mathbf{Q}$.

Proof It suffices to prove that

$$\sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} \geq \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t^*(\hat{x})} \tag{10.62}$$

for all $\mathbf{t} > 0$. The details are left as an exercise. Note in (10.61) that $\mathbf{t}^* > 0$ because $\mathbf{Q} > 0$. □

Since $I(\mathbf{p},\mathbf{Q})$ and $D(\mathbf{p},\mathbf{Q})$ are continuous in $\mathbf{Q}$, via an argument similar to the one we used in the proof of Theorem 10.2, we can replace the minimum over all $\mathbf{Q}$ in (10.57) by the infimum over all $\mathbf{Q} > 0$. By noting that the right hand side of (10.60) is equal to $I(\mathbf{p},\mathbf{Q})$ and

$$D(\mathbf{p},\mathbf{Q}) = \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}), \tag{10.63}$$

we can apply Lemma 10.3 to obtain

$$\begin{aligned} & R(D_s) - sD_s \\ &= \inf_{\mathbf{Q}>0} \left[\min_{\mathbf{t}>0} \sum_{x,\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} - s \sum_{x,\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}) \right] \qquad (10.64) \\ &= \inf_{\mathbf{Q}>0} \min_{\mathbf{t}>0} \left[\sum_{x,\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} - s \sum_{x,\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}) \right]. \qquad (10.65) \end{aligned}$$

Now in the double infimum in (10.14), let

$$\begin{aligned} f(\mathbf{Q},\mathbf{t}) &= \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} \\ &\quad - s \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}), \qquad (10.66) \end{aligned}$$

$$\begin{aligned} A_1 &= \Bigg\{ (Q(\hat{x}|x), (x,\hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}) : Q(\hat{x}|x) > 0, \\ &\qquad \sum_{\hat{x}} Q(\hat{x}|x) = 1 \text{ for all } x \in \mathcal{X} \Bigg\}, \qquad (10.67) \end{aligned}$$

and

$$A_2 = \{(t(\hat{x}), \hat{x} \in \hat{\mathcal{X}}) : t(\hat{x}) > 0 \text{ and } \textstyle\sum_{\hat{x}} t(\hat{x}) = 1\}, \qquad (10.68)$$

with $\mathbf{Q}$ and $\mathbf{t}$ playing the roles of $\mathbf{u}_1$ and $\mathbf{u}_2$, respectively. Then A_1 is a subset of $\Re^{|\mathcal{X}||\hat{\mathcal{X}}|}$ and A_2 is a subset of $\Re^{|\hat{\mathcal{X}}|}$, and it is readily checked that both A_1 and A_2 are convex. Since $s \leq 0$,

$$\begin{aligned} & f(\mathbf{Q},\mathbf{t}) \\ &= \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} - s \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}) \qquad (10.69) \\ &\geq \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t^*(\hat{x})} + 0 \qquad (10.70) \\ &= I(X;\hat{X}) \qquad (10.71) \\ &\geq 0. \qquad (10.72) \end{aligned}$$

Therefore, f is bounded from below.

The double infimum in (10.14) now becomes

$$\inf_{\mathbf{Q}\in A_1} \inf_{\mathbf{t}\in A_2} \left[\sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x) \log \frac{Q(\hat{x}|x)}{t(\hat{x})} - s \sum_x \sum_{\hat{x}} p(x)Q(\hat{x}|x)d(x,\hat{x}) \right], \qquad (10.73)$$

where the infimum over all $\mathbf{t} \in A_2$ is in fact a minimum. We then apply the alternating optimization algorithm described in Section 10.2 to compute f^*, the value of (10.73). First, we arbitrarily choose a *strictly positive* transition matrix in A_1 and let it be $\mathbf{Q}^{(0)}$. Then we define $\mathbf{t}^{(0)}$ and in general $\mathbf{t}^{(k)}$ for $k \geq 1$ by

$$t^{(k)}(\hat{x}) = \sum_x p(x)Q^{(k)}(\hat{x}|x) \tag{10.74}$$

in view of Lemma 10.3. In order to define $\mathbf{Q}^{(1)}$ and in general $\mathbf{Q}^{(k)}$ for $k \geq 1$, we need to find the $\mathbf{Q} \in A_1$ which minimizes f for a given $\mathbf{t} \in A_2$, where the constraints on $\mathbf{Q}$ are

$$Q(\hat{x}|x) > 0 \quad \text{for all } (x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}, \tag{10.75}$$

and

$$\sum_{\hat{x}} Q(\hat{x}|x) = 1 \quad \text{for all } x \in \mathcal{X}. \tag{10.76}$$

As we did for the computation of the channel capacity, we first ignore the positivity constraints in (10.75) when setting up the Lagrange multipliers. Then we obtain

$$Q(\hat{x}|x) = \frac{t(\hat{x})e^{sd(x,\hat{x})}}{\sum_{\hat{x}'} t(\hat{x}')e^{sd(x,\hat{x}')}} > 0. \tag{10.77}$$

The details are left as an exercise. We then define $\mathbf{Q}^{(k)}$ for $k \geq 1$ by

$$Q^{(k)}(\hat{x}|x) = \frac{t^{(k-1)}(\hat{x})e^{sd(x,\hat{x})}}{\sum_{\hat{x}'} t^{(k-1)}(\hat{x}')e^{sd(x,\hat{x}')}}. \tag{10.78}$$

It will be shown in the next section that $f^{(k)} = f(\mathbf{Q}^{(k)}, \mathbf{t}^{(k)}) \to f^*$ as $k \to \infty$. If there exists a unique point $(R(D_s), D_s)$ on the $R(D)$ curve such that the slope of a tangent at that point is equal to s, then

$$(I(\mathbf{p}, \mathbf{Q}^{(k)}), D(\mathbf{p}, \mathbf{Q}^{(k)})) \to (R(D_s), D_s). \tag{10.79}$$

Otherwise, $(I(\mathbf{p}, \mathbf{Q}^{(k)}), D(\mathbf{p}, \mathbf{Q}^{(k)}))$ is arbitrarily close to the segment of the $R(D)$ curve at which the slope is equal to s when k is sufficiently large. These facts are easily shown to be true.

10.3 CONVERGENCE

In this section, we first prove that if f is concave, then $f^{(k)} \to f^*$. We then apply this sufficient condition to prove the convergence of the BA algorithm for computing the channel capacity. The convergence of the BA algorithm for computing the rate distortion function can be proved likewise. The details are omitted.

10.3.1 A SUFFICIENT CONDITION

In the alternating optimization algorithm in Section 10.1, we see from (10.7) and (10.8) that

$$\mathbf{u}^{(k+1)} = (\mathbf{u}_1^{(k+1)}, \mathbf{u}_2^{(k+1)}) = (c_1(\mathbf{u}_2^{(k)}), c_2(c_1(\mathbf{u}_2^{(k)}))) \tag{10.80}$$

for $k \geq 0$. Define

$$\Delta f(\mathbf{u}) = f(c_1(\mathbf{u}_2), c_2(c_1(\mathbf{u}_2))) - f(\mathbf{u}_1, \mathbf{u}_2). \tag{10.81}$$

Then

$$f^{(k+1)} - f^{(k)} = f(\mathbf{u}^{(k+1)}) - f(\mathbf{u}^{(k)}) \tag{10.82}$$

$$= f(c_1(\mathbf{u}_2^{(k)}), c_2(c_1(\mathbf{u}_2^{(k)}))) - f(\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)}) \tag{10.83}$$

$$= \Delta f(\mathbf{u}^{(k)}). \tag{10.84}$$

We will prove that f being concave is sufficient for $f^{(k)} \to f^*$. To this end, we first prove that if f is concave, then the algorithm cannot be trapped at $\mathbf{u}$ if $f(\mathbf{u}) < f^*$.

LEMMA 10.4 *Let f be concave. If $f^{(k)} < f^*$, then $f^{(k+1)} > f^{(k)}$.*

Proof We will prove that $\Delta f(\mathbf{u}) > 0$ for any $\mathbf{u} \in A$ such that $f(\mathbf{u}) < f^*$. Then if $f^{(k)} = f(\mathbf{u}^{(k)}) < f^*$, we see from (10.84) that

$$f^{(k+1)} - f^{(k)} = \Delta f(\mathbf{u}^{(k)}) > 0, \tag{10.85}$$

and the lemma is proved.

Consider any $\mathbf{u} \in A$ such that $f(\mathbf{u}) < f^*$. We will prove by contradiction that $\Delta f(\mathbf{u}) > 0$. Assume $\Delta f(\mathbf{u}) = 0$. Then it follows from (10.81) that

$$f(c_1(\mathbf{u}_2), c_2(c_1(\mathbf{u}_2))) = f(\mathbf{u}_1, \mathbf{u}_2). \tag{10.86}$$

Now we see from (10.5) that

$$f(c_1(\mathbf{u}_2), c_2(c_1(\mathbf{u}_2))) \geq f(c_1(\mathbf{u}_2), \mathbf{u}_2). \tag{10.87}$$

If $c_1(\mathbf{u}_2) \neq \mathbf{u}_1$, then

$$f(c_1(\mathbf{u}_2), \mathbf{u}_2) > f(\mathbf{u}_1, \mathbf{u}_2) \tag{10.88}$$

because $c_1(\mathbf{u}_2)$ is unique. Combining (10.87) and (10.88), we have

$$f(c_1(\mathbf{u}_2), c_2(c_1(\mathbf{u}_2))) > f(\mathbf{u}_1, \mathbf{u}_2), \tag{10.89}$$

which is a contradiction to (10.86). Therefore,

$$\mathbf{u}_1 = c_1(\mathbf{u}_2). \tag{10.90}$$

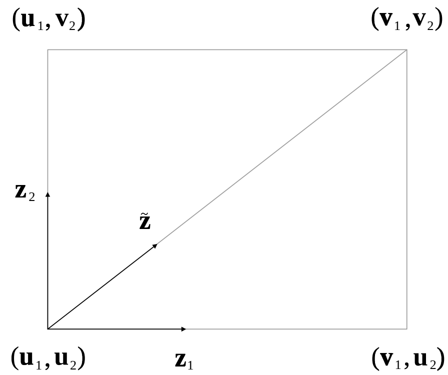

Figure 10.3. The vectors $\mathbf{u}$, $\mathbf{v}$, $\tilde{\mathbf{z}}$, $\mathbf{z}_1$, and $\mathbf{z}_2$.

Using this, we see from (10.86) that

$$f(\mathbf{u}_1, c_2(\mathbf{u}_1)) = f(\mathbf{u}_1, \mathbf{u}_2), \tag{10.91}$$

which implies

$$\mathbf{u}_2 = c_2(\mathbf{u}_1). \tag{10.92}$$

because $c_2(c_1(\mathbf{u}_2))$ is unique.

Since $f(\mathbf{u}) < f^*$, there exists $\mathbf{v} \in A$ such that

$$f(\mathbf{u}) < f(\mathbf{v}). \tag{10.93}$$

Consider

$$\mathbf{v} - \mathbf{u} = (\mathbf{v}_1 - \mathbf{u}_1, 0) + (0, \mathbf{v}_2 - \mathbf{u}_2). \tag{10.94}$$

Let $\tilde{\mathbf{z}}$ be the unit vector in the direction of $\mathbf{v} - \mathbf{u}$, $\mathbf{z}_1$ be the unit vector in the direction of $(\mathbf{v}_1 - \mathbf{u}_1, 0)$, and $\mathbf{z}_2$ be the unit vector in the direction of $(\mathbf{v}_2 - \mathbf{u}_2, 0)$. Then

$$\|\mathbf{v} - \mathbf{u}\|\tilde{\mathbf{z}} = \|\mathbf{v}_1 - \mathbf{u}_1\|\mathbf{z}_1 + \|\mathbf{v}_2 - \mathbf{u}_2\|\mathbf{z}_2, \tag{10.95}$$

or

$$\tilde{\mathbf{z}} = \alpha_1 \mathbf{z}_1 + \alpha_2 \mathbf{z}_2, \tag{10.96}$$

where

$$\alpha_i = \frac{\|\mathbf{v}_i - \mathbf{u}_i\|}{\|\mathbf{v} - \mathbf{u}\|}, \tag{10.97}$$

$i = 1, 2$. Figure 10.3 is an illustration of the vectors $\mathbf{u}$, $\mathbf{v}$, $\tilde{\mathbf{z}}$, $\mathbf{z}_1$, and $\mathbf{z}_2$.

We see from (10.90) that f attains its maximum value at $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ when $\mathbf{u}_2$ is fixed. In particular, f attains its maximum value at $\mathbf{u}$ alone the line passing through $(\mathbf{u}_1, \mathbf{u}_2)$ and $(\mathbf{v}_1, \mathbf{u}_2)$. Let $\bigtriangledown f$ denotes the gradient of

f. Since f is continuous and has continuous partial derivatives, the directional derivative of f at $\mathbf{u}$ in the direction of $\mathbf{z}_1$ exists and is given by $\nabla f \cdot \mathbf{z}_1$. It follows from the concavity of f that f is concave along the line passing through $(\mathbf{u}_1, \mathbf{u}_2)$ and $(\mathbf{v}_1, \mathbf{u}_2)$. Since f attains its maximum value at $\mathbf{u}$, the derivative of f along the line passing through $(\mathbf{u}_1, \mathbf{u}_2)$ and $(\mathbf{v}_1, \mathbf{u}_2)$ vanishes. Then we see that

$$\nabla f \cdot \mathbf{z}_1 = 0. \tag{10.98}$$

Similarly, we see from (10.92) that

$$\nabla f \cdot \mathbf{z}_2 = 0. \tag{10.99}$$

Then from (10.96), the directional derivative of f at $\mathbf{u}$ in the direction of $\tilde{\mathbf{z}}$ is given by

$$\nabla f \cdot \tilde{\mathbf{z}} = \alpha_1(\nabla f \cdot \mathbf{z}_1) + \alpha_2(\nabla f \cdot \mathbf{z}_2) = 0. \tag{10.100}$$

Since f is concave along the line passing through $\mathbf{u}$ and $\mathbf{v}$, this implies

$$f(\mathbf{u}) \geq f(\mathbf{v}), \tag{10.101}$$

which is a contradiction to (10.93). Hence, we conclude that $\Delta f(\mathbf{u}) > 0$. □

Although we have proved that the algorithm cannot be trapped at $\mathbf{u}$ if $f(\mathbf{u}) < f^*$, $f^{(k)}$ does not necessarily converge to f^* because the increment in $f^{(k)}$ in each step may be arbitrarily small. In order to prove the desired convergence, we will show in next theorem that this cannot be the case.

THEOREM 10.5 *If f is concave, then $f^{(k)} \to f^*$.*

Proof We have already shown in Section 10.1 that $f^{(k)}$ necessarily converges, say to f'. Hence, for any $\epsilon > 0$ and all sufficiently large k,

$$f' - \epsilon \leq f^{(k)} \leq f'. \tag{10.102}$$

Let

$$\gamma = \min_{\mathbf{u} \in A'} \Delta f(\mathbf{u}), \tag{10.103}$$

where

$$A' = \{\mathbf{u} \in A : f' - \epsilon \leq f(\mathbf{u}) \leq f'\}. \tag{10.104}$$

Since f has continuous partial derivatives, $\Delta f(\mathbf{u})$ is a continuous function of $\mathbf{u}$. Then the minimum in (10.103) exists because A' is compact[2].

[2] A' is compact because it is the inverse image of a closed interval under a continuous function and A is bounded.

We now show that $f' < f^*$ will lead to a contradiction if f is concave. If $f' < f^*$, then from Lemma 10.4, we see that $\Delta f(\mathbf{u}) > 0$ for all $\mathbf{u} \in A'$ and hence $\gamma > 0$. Since $f^{(k)} = f(\mathbf{u}^{(k)})$ satisfies (10.102), $\mathbf{u}^{(k)} \in A'$, and

$$f^{(k+1)} - f^{(k)} = \Delta f(\mathbf{u}^{(k)}) \geq \gamma \tag{10.105}$$

for all sufficiently large k. Therefore, no matter how smaller γ is, $f^{(k)}$ will eventually be greater than f', which is a contradiction to $f^{(k)} \to f'$. Hence, we conclude that $f^{(k)} \to f^*$. □

10.3.2 CONVERGENCE TO THE CHANNEL CAPACITY

In order to show that the BA algorithm for computing the channel capacity converges as intended, i.e., $f^{(k)} \to C$, we only need to show that the function f defined in (10.39) is concave. Toward this end, for

$$f(\mathbf{r}, \mathbf{q}) = \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} \tag{10.106}$$

defined in (10.39), we consider two ordered pairs $(\mathbf{r}_1, \mathbf{q}_1)$ and $(\mathbf{r}_2, \mathbf{q}_2)$ in A, where A_1 and A_2 are defined in (10.40) and (10.41), respectively. For any $0 \leq \lambda \leq 1$ and $\bar{\lambda} = 1 - \lambda$, an application of the log-sum inequality (Theorem 2.31) gives

$$\begin{aligned}
&(\lambda r_1(x) + \bar{\lambda} r_2(x)) \log \frac{\lambda r_1(x) + \bar{\lambda} r_2(x)}{\lambda q_1(x|y) + \bar{\lambda} q_2(x|y)} \\
&\leq \quad \lambda r_1(x) \log \frac{r_1(x)}{q_1(x|y)} + \bar{\lambda} r_2(x) \log \frac{r_2(x)}{q_2(x|y)}.
\end{aligned} \tag{10.107}$$

Taking reciprocal in the logarithms yields

$$\begin{aligned}
&(\lambda r_1(x) + \bar{\lambda} r_2(x)) \log \frac{\lambda q_1(x|y) + \bar{\lambda} q_2(x|y)}{\lambda r_1(x) + \bar{\lambda} r_2(x)} \\
&\geq \quad \lambda r_1(x) \log \frac{q_1(x|y)}{r_1(x)} + \bar{\lambda} r_2(x) \log \frac{q_2(x|y)}{r_2(x)},
\end{aligned} \tag{10.108}$$

and upon multiplying by $p(y|x)$ and summing over all x and y, we obtain

$$f(\lambda \mathbf{r}_1 + \bar{\lambda} \mathbf{r}_2, \lambda \mathbf{q}_1 + \bar{\lambda} \mathbf{q}_2) \geq \lambda f(\mathbf{r}_1, \mathbf{q}_1) + \bar{\lambda} f(\mathbf{r}_2, \mathbf{q}_2). \tag{10.109}$$

Therefore, f is concave. Hence, we have shown that $f^{(k)} \to C$.

PROBLEMS

1. Implement the BA algorithm for computing channel capacity.

2. Implement the BA algorithm for computing the rate-distortion function.

3. Explain why in the BA Algorithm for computing channel capacity, we should not choose an initial input distribution which contains zero probability masses.

4. Prove Lemma 10.3.

5. Consider $f(\mathbf{Q}, \mathbf{t})$ in the BA algorithm for computing the rate-distortion function.

   a) Show that for fixed s and $\mathbf{t}$, $f(\mathbf{Q}, \mathbf{t})$ is minimized by

   $$Q(\hat{x}|x) = \frac{t(\hat{x})e^{sd(x,\hat{x})}}{\sum_{\hat{x}'} t(\hat{x}')e^{sd(x,\hat{x}')}}.$$

   b) Show that $f(\mathbf{Q}, \mathbf{t})$ is convex.

HISTORICAL NOTES

An iterative algorithm for computing the channel capacity was developed by Arimoto [14], where the convergence of the algorithm was proved. Blahut [27] independently developed two similar algorithms, the first for computing the channel capacity and the second for computing the rate distortion function. The convergence of Blahut's second algorithm was proved by Csiszár [51]. These two algorithms are now commonly referred to as the Blahut-Arimoto algorithms. The simplified proof of convergence in this chapter is based on Yeung and Berger [217].

The Blahut-Arimoto algorithms are special cases of a general iterative algorithm due to Csiszár and Tusnády [55] which also include the EM algorithm [59] for fitting models from incomplete data and the algorithm for finding the log-optimal portfolio for a stock market due to Cover [46].

Bibliography

[1] J. Abrahams, "Code and parse trees for lossless source encoding," *Comm. Inform. & Syst.*, 1: 113-146, 2001.

[2] N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York, 1963.

[3] Y. S. Abu-Mostafa, Ed., *Complexity in Information Theory*, Springer-Verlag, New York, 1988.

[4] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterizations*, Academic Press, New York, 1975.

[5] R. Ahlswede, B. Balkenhol and L. Khachatrian, "Some properties of fix-free codes," preprint 97-039, Sonderforschungsbereich 343, Universität Bielefeld, 1997.

[6] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inform. Theory*, IT-46: 1204-1216, 2000.

[7] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inform. Theory*, IT-21: 629-637, 1975.

[8] R. Ahlswede and I. Wegener, *Suchprobleme*, Teubner Studienbcher. B. G. Teubner, Stuttgart, 1979 (in German). English translation: *Search Problems*, Wiley, New York, 1987.

[9] R. Ahlswede and J. Wolfowitz, "The capacity of a channel with arbitrarily varying cpf's and binary output alphabet," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 15: 186-194, 1970.

[10] P. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Ann. Prob.*, 16: 899-909, 1988.

[11] S. Amari, *Differential-Geometrical Methods in Statistics*, Springer-Verlag, New York, 1985.

[12] J. B. Anderson and S. Mohan, *Source and Channel Coding: An Algorithmic Approach*, Kluwer Academic Publishers, Boston, 1991.

[13] S. Arimoto, "Encoding and decoding of p-ary group codes and the correction system," *Information Processing in Japan*, 2: 321-325, 1961 (in Japanese).

[14] S. Arimoto, "An algorithm for calculating the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, IT-18: 14-20, 1972.

[15] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, IT-19: 357-359, 1973.

[16] R. B. Ash, *Information Theory*, Interscience, New York, 1965.

[17] E. Ayanoglu, R. D. Gitlin, C.-L. I, and J. Mazo, "Diversity coding for transparent self-healing and fault-tolerant communication networks," 1990 IEEE International Symposium on Information Theory, San Diego, CA, Jan. 1990.

[18] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem," *Ann. Prob.*, 13: 1292-1303, 1985.

[19] L. A. Bassalygo, R. L. Dobrushin, and M. S. Pinsker, "Kolmogorov remembered," *IEEE Trans. Inform. Theory*, IT-34: 174-175, 1988.

[20] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[21] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed., CISM Courses and Lectures #229, Springer-Verlag, New York, 1978.

[22] T. Berger and R. W. Yeung, "Multiterminal source coding with encoder breakdown," *IEEE Trans. Inform. Theory*, IT-35: 237-244, 1989.

[23] E. R. Berlekamp, "Block coding for the binary symmetric channel with noiseless, delayless feedback," in H. B. Mann, *Error Correcting Codes*, Wiley, New York, 1968.

[24] E. R. Berlekamp, Ed., *Key Papers in the Development of Coding Theory*, IEEE Press, New York, 1974.

[25] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," Proceedings of the 1993 International Conferences on Communications, 1064-1070, 1993.

[26] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Stat.*, 31: 558-567, 1960.

[27] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. Inform. Theory*, IT-18: 460-473, 1972.

[28] R. E. Blahut, "Information bounds of the Fano-Kullback type," *IEEE Trans. Inform. Theory*, IT-22: 410-421, 1976.

[29] R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, Massachusetts, 1983.

[30] R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, Massachusetts, 1987.

[31] R. E. Blahut, D. J. Costello, Jr., U. Maurer, and T. Mittelholzer, Ed., *Communications and Cryptography: Two Sides of One Tapestry*, Kluwer Academic Publishers, Boston, 1994.

[32] C. Blundo, A. De Santis, R. De Simone, and U. Vaccaro, "Tight bounds on the information rate of secret sharing schemes," *Designs, Codes and Cryptography*, 11: 107-110, 1997.

[33] R. C. Bose and D. K. Ray-Chaudhuri, "On a class of error correcting binary group codes," *Inform. Contr.*, 3: 68-79, Mar. 1960.

[34] L. Breiman, "The individual ergodic theorems of information theory," *Ann. Math. Stat.*, 28: 809-811, 1957.

[35] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Technical Report 124, Digital Equipment Corporation, 1994.

[36] R. Calderbank and N. J. A. Sloane, "Obituary: Claude Shannon (1916-2001)," *Nature*, 410: 768, April 12, 2001.

[37] R. M. Capocelli, A. De Santis, L. Gargano, and U. Vaccaro, "On the size of shares for secret sharing schemes," *J. Cryptology*, 6: 157-168, 1993.

[38] H. L. Chan (T. H. Chan), "Aspects of information inequalities and its applications," M.Phil. thesis, The Chinese University of Hong Kong, Jun. 1998.

[39] T. H. Chan, "A combinatorial approach to information inequalities," to appear in *Comm. Inform. & Syst.*.

[40] T. H. Chan and R. W. Yeung, "On a relation between information inequalities and group theory," to appear in *IEEE Trans. Inform. Theory*.

[41] T. H. Chan and R. W. Yeung, "Factorization of positive functions," in preparation.

[42] G. J. Chatin, *Algorithmic Information Theory*, Cambridge Univ. Press, Cambridge, 1987.

[43] H. Chernoff, "A measure of the asymptotic efficiency of test of a hypothesis based on a sum of observations," *Ann. Math. Stat.*, 23: 493-507, 1952.

[44] K. L. Chung, "A note on the ergodic theorem of information theory," *Ann. Math. Stat.*, 32: 612-614, 1961.

[45] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, IT-21: 226-228, 1975.

[46] T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory*, IT-30: 369-373, 1984.

[47] T. M. Cover, P. Gắcs, and R. M. Gray, "Kolmogorov's contribution to information theory and algorithmic complexity," *Ann. Prob.*, 17: 840-865, 1989.

[48] T. M. Cover and S. K. Leung, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, IT-24: 331-338, 1978.

[49] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[50] I. Csiszár, “Information type measures of difference of probability distributions and indirect observations,” *Studia Sci. Math. Hungar.*, 2: 229-318, 1967.

[51] I. Csiszár, “On the computation of rate-distortion functions,” *IEEE Trans. Inform. Theory*, IT-20: 122-124, 1974.

[52] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[53] I. Csiszár and P. Narayan, “Arbitrarily varying channels with constrained inputs and states,” *IEEE Trans. Inform. Theory*, IT-34: 27-34, 1988.

[54] I. Csiszár and P. Narayan, “The capacity of the arbitrarily varying channel revisited: Positivity, constraints,” *IEEE Trans. Inform. Theory*, IT-34: 181-193, 1988.

[55] I. Csiszár and G. Tusnády, “Information geometry and alternating minimization procedures,” *Statistics and Decisions*, Supplement Issue 1: 205-237, 1984.

[56] G. B. Dantzig, *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, New Jersey, 1962.

[57] L. D. Davisson, “Universal noiseless coding,” *IEEE Trans. Inform. Theory*, IT-19: 783-795, 1973.

[58] A. P. Dawid, “Conditional independence in statistical theory (with discussion)," *J. Roy. Statist. Soc., Series B*, 41: 1-31, 1979.

[59] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood form incomplete data via the EM algorithm,” *Journal Royal Stat. Soc., Series B*, 39: 1-38, 1977.

[60] G. Dueck and J. Körner, “Reliability function of a discrete memoryless channel at rates above capacity,” *IEEE Trans. Inform. Theory*, IT-25: 82-85, 1979.

[61] P. Elias, “Universal codeword sets and representations of the integers,” *IEEE Trans. Inform. Theory*, IT-21: 194-203, 1975.

[62] *Encyclopedia Britanica*, http://www/britanica.com/.

[63] R. M. Fano, Class notes for Transmission of Information, Course 6.574, MIT, Cambridge, Massachusetts, 1952.

[64] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*, Wiley, New York, 1961.

[65] A. Feinstein, “A new basic theorem of information theory,” *IRE Trans. Inform. Theory*, IT-4: 2-22, 1954.

[66] A. Feinstein, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[67] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, New York, 1950.

[68] B. M. Fitingof, "Coding in the case of unknown and changing message statistics," *PPI* 2: 3-11, 1966 (in Russian).

[69] L. K. Ford, Jr. and D. K. Fulkerson, *Flows in Networks*, Princeton Univ. Press, Princeton, New Jersey, 1962.

[70] G. D. Forney, Jr., "Convolutional codes I: Algebraic structure," *IEEE Trans. Inform. Theory*, IT-16: 720 - 738, 1970.

[71] G. D. Forney, Jr., Information Theory, unpublished course notes, Stanford University, 1972.

[72] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, 61: 268-278, 1973.

[73] F. Fu, R. W. Yeung, and R. Zamir, "On the rate-distortion region for multiple descriptions," submitted to *IEEE Trans. Inform. Theory*.

[74] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Inform. Contr.*, 39: 55-72, 1978.

[75] R. G. Gallager, "Low-density parity-check codes," *IEEE Trans. Inform. Theory*, IT-8: 21-28, Jan. 1962.

[76] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, IT-11: 3-18, 1965.

[77] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[78] Y. Ge and Z. Ye, "Information-theoretic characterizations of lattice conditional independence models," submitted to *Ann. Stat.*

[79] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston, 1992.

[80] S. Goldman, *Information Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1953.

[81] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[82] S. Guiasu, *Information Theory with Applications*, McGraw-Hill, New York, 1976.

[83] B. E. Hajek and T. Berger, "A decomposition theorem for binary Markov random fields," *Ann. Prob.*, 15: 1112-1125, 1987.

[84] D. Hammer, A. Romashchenko, A. Shen, and N. K. Vereshchagin, *J. Comp. & Syst. Sci.*, 60: 442-464, 2000.

[85] R. V. Hamming, "Error detecting and error correcting codes," *Bell Sys. Tech. Journal*, 29: 147-160, 1950.

[86] T. S. Han, "Linear dependence structure of the entropy space," *Inform. Contr.*, 29: 337-368, 1975.

[87] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inform. Contr.*, 36: 133-156, 1978.

[88] T. S. Han, "A uniqueness of Shannon's information distance and related non-negativity problems," *J. Comb., Inform., & Syst. Sci.*, 6: 320-321, 1981.

[89] T. S. Han, "An information-spectrum approach to source coding theorems with a fidelity criterion," *IEEE Trans. Inform. Theory*, IT-43: 1145-1164, 1997.

[90] T. S. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Trans. Inform. Theory*, IT-26: 277-288, 1980.

[91] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, 2nd ed., Cambridge Univ. Press, London, 1952.

[92] K. P. Hau, "Multilevel diversity coding with independent data streams," M.Phil. thesis, The Chinese University of Hong Kong, Jun. 1995.

[93] C. Heegard and S. B. Wicker, *Turbo Coding*, Kluwer Academic Publishers, Boston, 1999.

[94] A. Hocquenghem, "Codes correcteurs d'erreurs," *Chiffres*, 2: 147-156, 1959.

[95] Y. Horibe, "An improved bound for weight-balanced tree," *Inform. Contr.*, 34: 148-151, 1977.

[96] Hu Guoding, "On the amount of Information," *Teor. Veroyatnost. i Primenen.*, 4: 447-455, 1962 (in Russian).

[97] D. A. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, 40: 1098-1101, 1952.

[98] L. P. Hyvarinen, *Information Theory for Systems Engineers*, Springer-Verlag, Berlin, 1968.

[99] A. W. Ingleton, "Representation of matroids," in *Combinatorial Mathematics and Its Applications*, D. J. A. Welsh, Ed., 149-167, Academic Press, London, 1971.

[100] E. T. Jaynes, "On the rationale of maximum entropy methods," *Proc. IEEE*, 70: 939-052, 1982.

[101] F. Jelinek, *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.

[102] V. D. Jerohin, "ϵ-entropy of discrete random objects," *Teor. Veroyatnost. i Primenen*, 3: 103-107, 1958.

[103] O. Johnsen, "On the redundancy of binary Huffman codes," *IEEE Trans. Inform. Theory*, IT-26: 220-222, 1980.

[104] G. A. Jones and J. M. Jones, *Information and Coding Theory*, Springer, London, 2000.

[105] Y. Kakihara, *Abstract Methods in Information Theory*, World-Scientific, Singapore, 1999.

[106] J. Karush, "A simple proof of an inequality of McMillan," *IRE Trans. Inform. Theory*, 7: 118, 1961.

[107] T. Kawabata, "Gaussian multiterminal source coding," Master thesis, Math. Eng., Univ. of Tokyo, Japan, Feb. 1980.

[108] T. Kawabata and R. W. Yeung, "The structure of the I-Measure of a Markov chain," *IEEE Trans. Inform. Theory,* IT-38: 1146-1149, 1992.

[109] A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover, New York, 1957.

[110] J. C. Kieffer, E.-h. Yang, "Grammar-based codes: A new class of universal lossless source codes," *IEEE Trans. Inform. Theory*, IT-46: 737-754, 2000.

[111] R. Kindermann and J. Snell, *Markov Random Fields and Their Applications*, American Math. Soc., Providence, Rhode Island, 1980.

[112] R. Koetter and M. Médard, "An algebraic approach to network coding," 2001 IEEE International Symposium on Information Theory, Washington, D.C., Jun. 2001.

[113] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IEEE Trans. Inform. Theory*, IT-2: 102-108, 1956.

[114] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, 1: 4-7, 1965.

[115] A. N. Kolmogorov, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, IT-14: 662-664, 1968.

[116] L. G. Kraft, "A device for quantizing, grouping and coding amplitude modulated pulses," M.S. thesis, Dept. of E.E., MIT, 1949.

[117] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[118] S. Kullback, *Topics in Statistical Information Theory*, Springer-Verlag, Berlin, 1987.

[119] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, 22: 79-86, 1951.

[120] G. G. Langdon, "An introduction to arithmetic coding," *IBM J. Res. Devel.*, 28: 135-149, 1984.

[121] S. L. Lauritzen, *Graphical Models*, Oxford Science Publications, Oxford, 1996.

[122] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed., Springer, New York, 1997.

[123] S.-Y. R. Li, R. W. Yeung and N. Cai, "Linear network coding," to appear in *IEEE Trans. Inform. Theory*.

[124] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[125] T. Linder, V. Tarokh, and K. Zeger, "Existence of optimal codes for infinite source alphabets," *IEEE Trans. Inform. Theory*, IT-43: 2026-2028, 1997.

[126] L. Lovasz, "On the Shannon capacity of a graph," *IEEE Trans. Inform. Theory*, IT-25: 1-7, 1979.

[127] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, IT-45: 399-431, Mar. 1999.

[128] F. M. Malvestuto, "A unique formal system for binary decompositions of database relations, probability distributions, and graphs," *Inform. Sci.*, 59: 21-52, 1992; with Comment by F. M. Malvestuto and M. Studený, *Inform. Sci.*, 63: 1-2, 1992.

[129] M. Mansuripur, *Introduction to Information Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[130] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, IT-20: 197 - 199, 1974.

[131] J. L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Trans. Inform. Theory*, IT-15: 122-127, 1969.

[132] J. L. Massey, "Causality, feedback and directed information," in *Proc. 1990 Int. Symp. on Inform. Theory and Its Applications*, 303-305, 1990.

[133] J. L. Massey, "Contemporary cryptology: An introduction," in *Contemporary Cryptology: The Science of Information Integrity*, G. J. Simmons, Ed., IEEE Press, Piscataway, New Jersey, 1992.

[134] A. M. Mathai and P. N. Rathie, *Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications*, Wiley, New York, 1975.

[135] F. Matúš, "Probabilistic conditional independence structures and matroid theory: Background," *Int. J. of General Syst.*, 22: 185-196, 1994.

[136] F. Matúš, "Conditional independences among four random variables II," *Combinatorics, Probability & Computing*, 4: 407-417, 1995.

[137] F. Matúš, "Conditional independences among four random variables III: Final conclusion," *Combinatorics, Probability & Computing*, 8: 269-276, 1999.

[138] F. Matúš and M. Studený, "Conditional independences among four random variables I," *Combinatorics, Probability & Computing*, 4: 269-278, 1995.

[139] R. J. McEliece, *The Theory of Information and Coding*, Addison-Wesley, Reading, Massachusetts, 1977.

[140] W. J. McGill, "Multivariate information transmission," *Transactions PGIT, 1954 Symposium on Information Theory*, PGIT-4: pp. 93-111, 1954.

[141] B. McMillan, "The basic theorems of information theory," *Ann. Math. Stat.*, 24: 196-219, 1953.

[142] B. McMillan, "Two inequalities implied by unique decipherability," *IRE Trans. Inform. Theory*, 2: 115-116, 1956.

[143] S. C. Moy, "Generalization of the Shannon-McMillan theorem," *Pacific J. Math.*, 11: 705-714, 1961.

[144] J. K. Omura, "A coding theorem for discrete-time sources," *IEEE Trans. Inform. Theory*, IT-19: 490-498, 1973.

[145] J. M. Ooi, *Coding for Channels with Feedback*, Kluwer Academic Publishers, Boston, 1998.

[146] A. Orlitsky, “Worst-case interactive communication I: Two messages are almost optimal,” *IEEE Trans. Inform. Theory*, IT-36: 1111-1126, 1990.

[147] A. Orlitsky, “Worst-case interactive communication—II: Two messages are not optimal,” *IEEE Trans. Inform. Theory*, IT-37: 995-1005, 1991.

[148] D. S. Ornstein, “Bernoulli shifts with the same entropy are isomorphic,” *Advances in Math.*, 4: 337-352, 1970.

[149] J. G. Oxley, *Matroid Theory*, Oxford Univ. Press, Oxford, 1992.

[150] A. Papoulis, *Probability, Random Variables and Stochastic Processes,* 2nd ed., McGraw-Hill, New York, 1984.

[151] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman, San Meteo, California, 1988.

[152] A. Perez, “Extensions of Shannon-McMillan’s limit theorem to more general stochastic processes,” in Trans. Third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, 545-574, Prague, 1964.

[153] J. R. Pierce, *An Introduction to Information Theory : Symbols, Signals and Noise*, 2nd rev. ed., Dover, New York, 1980.

[154] J. T. Pinkston, “An application of rate-distortion theory to a converse to the coding theorem,” *IEEE Trans. Inform. Theory*, IT-15: 66-71, 1969.

[155] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Vol. 7 of the series *Problemy Peredači Informacii*, AN SSSR, Moscow, 1960 (in Russian). English translation: Holden-Day, San Francisco, 1964.

[156] N. Pippenger, “What are the laws of information theory?” 1986 Special Problems on Communication and Computation Conference, Palo Alto, California, Sept. 3-5, 1986.

[157] C. Preston, *Random Fields*, Springer-Verlag, New York, 1974.

[158] M. O. Rabin, “Efficient dispersal of information for security, load balancing, and fault-tolerance,” *J. ACM*, 36: 335-348, 1989.

[159] I. S. Reed and G. Solomon, “Polynomial codes over certain finite fields,” *SIAM Journal Appl. Math.*, 8: 300-304, 1960.

[160] A. Rényi, *Foundations of Probability*, Holden-Day, San Francisco, 1970.

[161] F. M. Reza, *An Introduction to Information Theory*, McGraw-Hill, New York, 1961.

[162] J. Rissanen, “Generalized Kraft inequality and arithmetic coding,” *IBM J. Res. Devel.*, 20: 198, 1976.

[163] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Trans. Inform. Theory*, IT-30: 629-636, 1984.

[164] J. R. Roche, “Distributed information storage,” Ph.D. thesis, Stanford University, Mar. 1992.

[165] J. R. Roche, A. Dembo, and A. Nobel, “Distributed information storage,” 1988 IEEE International Symposium on Information Theory, Kobe, Japan, Jun. 1988.

[166] J. R. Roche, R. W. Yeung, and K. P. Hau, “Symmetrical multilevel diversity coding,” *IEEE Trans. Inform. Theory*, IT-43: 1059-1064, 1997.

[167] R. T. Rockafellar, *Convex Analysis*, Princeton Univ. Press, Princeton, New Jersey, 1970.

[168] A. Romashchenko, A. Shen, and N. K. Vereshchagin, “Combinatorial interpretation of Kolmogorov complexity,” *Electronic Colloquium on Computational Complexity*, vol.7, 2000.

[169] K. Rose, “A mapping approach to rate-distortion computation and analysis,” *IEEE Trans. Inform. Theory*, IT-40: 1939-1952, 1994.

[170] S. Shamai, S. Verdú, “The empirical distribution of good codes,” *IEEE Trans. Inform. Theory*, IT-43: 836-846, 1997.

[171] S. Shamai, S. Verdú, and R. Zamir, “Systematic lossy source/channel coding,” *IEEE Trans. Inform. Theory*, IT-44: 564-579, 1998.

[172] A. Shamir, “How to share a secret,” *Comm. ACM*, 22: 612-613, 1979.

[173] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.

[174] C. E. Shannon, “Communication theory of secrecy systems,” *Bell Sys. Tech. Journal*, 28: 656-715, 1949.

[175] C. E. Shannon, “The zero-error capacity of a noisy channel,” *IRE Trans. Inform. Theory*, IT-2: 8-19, 1956.

[176] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE National Convention Record, Part 4*, 142-163, 1959.

[177] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding in discrete memoryless channels,” *Inform. Contr.*, 10: 65-103 (Part I), 522-552 (Part II), 1967.

[178] C. E. Shannon and W. W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, Illinois, 1949.

[179] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, American Math. Soc., Providence, Rhode Island, 1996.

[180] J. E. Shore and R. W. Johnson, “Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,” *IEEE Trans. Inform. Theory*, IT-26: 26-37, 1980.

[181] I. Shunsuke, *Information theory for continuous systems*, World Scientific, Singapore, 1993.

[182] M. Simonnard, *Linear Programming*, translated by William S. Jewell, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

[183] D. S. Slepian, Ed., *Key Papers in the Development of Information Theory*, IEEE Press, New York, 1974.

[184] D. S. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, IT-19: 471-480, 1973.

[185] N. J. A. Sloane and A. D. Wyner, Ed., *Claude Elwood Shannon Collected Papers*, IEEE Press, New York, 1993.

[186] L. Song, R. W. Yeung, and N. Cai, "A separation theorem for point-to-point communication networks," submitted to *IEEE Trans. Inform. Theory*.

[187] L. Song, R. W. Yeung and N. Cai, "Zero-error network coding for acyclic networks," submitted to *IEEE Trans. Inform. Theory*.

[188] F. Spitzer, "Random fields and interacting particle systems," M. A. A. Summer Seminar Notes, 1971.

[189] M. Studený, "Multiinformation and the problem of characterization of conditional-independence relations," *Problems Control Inform. Theory*, 18: 1, 3-16, 1989.

[190] J. C. A. van der Lubbe, *Information Theory*, Cambridge Univ. Press, Cambridge, 1997 (English translation).

[191] E. C. van der Meulen, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Trans. Inform. Theory*, IT-23: 1-37, 1977.

[192] E. C. van der Meulen, "Some reflections on the interference channel," in *Communications and Cryptography: Two Side of One Tapestry*, R. E. Blahut, D. J. Costello, Jr., U. Maurer, and T. Mittelholzer, Ed., Kluwer Academic Publishers, Boston, 1994.

[193] M. van Dijk, "On the information rate of perfect secret sharing schemes," *Designs, Codes and Cryptography*, 6: 143-169, 1995.

[194] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, IT-41: 44-54, 1995.

[195] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, IT-40: 1147-1157, 1994.

[196] S. Verdú and T. S. Han, "The role of the asymptotic equipartition property in noiseless source coding," *IEEE Trans. Inform. Theory*, IT-43: 847-857, 1997.

[197] S. Verdú and S. W. McLaughlin, Ed., *Information Theory : 50 years of Discovery*, IEEE Press, New York, 2000.

[198] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, IT-13: 260-269, 1967.

[199] A. J. Viterbi and J. K. Omura, *Principles of Digital Communications and Coding*, McGraw-Hill, New York, 1979.

[200] T. A. Welch, “A technique for high-performance data compression,” *Computer*, 17: 8-19, 1984.

[201] P. M. Woodard, *Probability and Information Theory with Applications to Radar*, McGraw-Hill, New York, 1953.

[202] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, New Jersey, 1995.

[203] S. B. Wicker and V. K. Bhargava, Ed., *Reed-Solomon Codes and Their Applications*, IEEE Press, Piscataway, New Jersey, 1994.

[204] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context-tree weighting method: basic properties,” *IEEE Trans. Inform. Theory*, IT-41: 653-664, 1995.

[205] J. Wolfowitz, “The coding of messages subject to chance errors,” *Illinois Journal of Mathematics*, 1: 591-606, 1957.

[206] J. Wolfowitz, *Coding Theorems of Information Theory*, Springer, Berlin-Heidelberg, 2nd ed., 1964, 3rd ed., 1978.

[207] A. D. Wyner, “On source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, IT-21: 294-300, 1975.

[208] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, IT-22: 1-10, 1976.

[209] E.-h. Yang and J. C. Kieffer, “Efficient universal lossless data compression algorithms based on a greedy sequential grammar transform – Part one: Without context models,” *IEEE Trans. Inform. Theory*, IT-46: 755-777, 2000.

[210] C. Ye and R. W. Yeung, “Some basic properties of fix-free codes,” *IEEE Trans. Inform. Theory*, IT-47: 72-87, 2001.

[211] C. Ye and R. W. Yeung, “A simple upper bound on the redundancy of Huffman codes,” to appear in *IEEE Trans. Inform. Theory*.

[212] Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*, Science Press, Beijing/New York, 1998.

[213] R. W. Yeung, “A new outlook on Shannon’s information measures,” *IEEE Trans. Inform. Theory*, IT-37: 466-474, 1991.

[214] R. W. Yeung, “Local redundancy and progressive bounds on the redundancy of a Huffman code," *IEEE Trans. Inform. Theory*, IT-37: 687-691, 1991.

[215] R. W. Yeung, “Multilevel diversity coding with distortion,” *IEEE Trans. Inform. Theory*, IT-41: 412-422, 1995.

[216] R. W. Yeung, “A framework for linear information inequalities,” *IEEE Trans. Inform. Theory*, IT-43: 1924-1934, 1997.

[217] R. W. Yeung and T. Berger, “Multi-way alternating minimization,” 1995 IEEE Internation Symposium on Information Theory, Whistler, British Columbia, Canada, Sept. 1995.

[218] R. W. Yeung, T. T. Lee and Z. Ye, "Information-theoretic characterization of conditional mutual independence and Markov random fields," to appear in *IEEE Trans. Inform. Theory*.

[219] R. W. Yeung and Z. Zhang, "On symmetrical multilevel diversity coding," *IEEE Trans. Inform. Theory*, IT-45: 609-621, 1999.

[220] R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Inform. Theory*, IT-45: 1111-1120, 1999.

[221] R. W. Yeung and Z. Zhang, "A class of non-Shannon-type information inequalities and their applications," *Comm. Inform. & Syst.*, 1: 87-100, 2001.

[222] Z. Zhang and R. W. Yeung, "A non-Shannon-type conditional inequality of information quantities," *IEEE Trans. Inform. Theory*, IT-43: 1982-1986, 1997.

[223] Z. Zhang and R. W. Yeung, "On characterization of entropy function via information inequalities," *IEEE Trans. Inform. Theory*, IT-44: 1440-1452, 1998.

[224] S. Zimmerman, "An optimal search procedure," *Am. Math. Monthly*, 66: 8, 690-693, 1959.

[225] K. Sh. Zigangirov, "Number of correctable errors for transmission over a binary symmetrical channel with feedback," *Problems Inform. Transmission*, 12: 85-97, 1976. Translated from *Problemi Peredachi Informatsii*, 12: 3-19 (in Russian).

[226] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, IT-23: 337-343, 1977.

[227] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, IT-24: 530-536, 1978.

Index