Elements of Information Theory
Thomas M. Cover, Joy A. Thomas
Copyright © 1991 John Wiley & Sons, Inc.
Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 16

Inequalities in Information Theory

This chapter summarizes and reorganizes the inequalities found throughout this book. A number of new inequalities on the entropy rates of subsets and the relationship of entropy and $\mathcal{L}_p$ norms are also developed. The intimate relationship between Fisher information and entropy is explored, culminating in a common proof of the entropy power inequality and the Brunn-Minkowski inequality. We also explore the parallels between the inequalities in information theory and inequalities in other branches of mathematics such as matrix theory and probability theory.

16.1 BASIC INEQUALITIES OF INFORMATION THEORY

Many of the basic inequalities of information theory follow directly from convexity.

Definition: A function f is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2) \tag{16.1}$$

for all $0 \le \lambda \le 1$ and all x_1 and x_2 in the convex domain of f.

Theorem 16.1.1 (*Theorem 2.6.2: Jensen's inequality*): *If f is convex, then*

$$f(EX) \le Ef(X)\,. \tag{16.2}$$

Lemma 16.1.1: *The function* $\log x$ *is a concave function and* $x \log x$ *is a convex function of* x, *for* $0 \le x < \infty$.

Theorem 16.1.2 (*Theorem 2.7.1: Log sum inequality*): *For positive numbers,* $a_1, a_2, \dots, a_n$ *and* $b_1, b_2, \dots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \ge \left(\sum_{i=1}^{n} a_i\right) \log \frac{(\Sigma_{i=1}^{n} a_i)}{(\Sigma_{i=1}^{n} b_i)} \tag{16.3}$$

with equality iff $\frac{a_i}{b_i}$ = *constant.*

We have the following properties of entropy from Section 2.1.

Definition: The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathscr{X}} p(x) \log p(x) . \tag{16.4}$$

Theorem 16.1.3 (*Lemma 2.1.1, Theorem 2.6.4: Entropy bound*):

$$0 \le H(X) \le \log|\mathscr{X}| \tag{16.5}$$

Theorem 16.1.4 (*Theorem 2.6.5: Conditioning reduces entropy*): *For any two random variables* X and Y,

$$H(X|Y) \le H(X) , \tag{16.6}$$

with equality iff X and Y are independent.

Theorem 16.1.5 (*Theorem 2.5.1 with Theorem 2.6.6*: *Chain rule*):

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) \le \sum_{i=1}^{n} H(X_i) , \tag{16.7}$$

with equality iff $X_1, X_2, \dots, X_n$ *are independent.*

Theorem 16.1.6 (*Theorem 2.7.3*): $H(\mathbf{p})$ *is a concave function of* $\mathbf{p}$.

We now state some properties of relative entropy and mutual information (Section 2.3).

Definition: The *relative entropy* or *Kullback Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ on the same set $\mathscr{X}$ is defined by

$$D(p\|q)=\sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\,. \tag{16.8}$$

Definition: The mutual information between two random variables X and Y is defined by

$$I(X;Y)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=D(p(x,y)\|p(x)p(y))\,. \tag{16.9}$$

The following basic information inequality can be used to prove many of the other inequalities in this chapter.

Theorem 16.1.7 (*Theorem 2.6.3: Information inequality*): *For any two probability mass functions* $\mathbf{p}$ *and* $\mathbf{q}$,

$$D(\mathbf{p}\|\mathbf{q})\geq 0 \tag{16.10}$$

with equality iff $p(x)=q(x)$ *for all* $x\in\mathcal{X}$.

Corollary: *For any two random variables, X and Y,*

$$I(X;Y)=D(p(x,y)\|p(x)p(y))\geq 0 \tag{16.11}$$

with equality iff $p(x,y)=p(x)p(y)$, i.e., X and Y are independent.

Theorem 16.1.8 (*Theorem 2.7.2: Convexity of relative entropy*): $D(p\|q)$ is convex in the pair (p,q).

Theorem 16.1.9 (*Section 2.4*):

$$I(X;Y)=H(X)-H(X|Y)\,, \tag{16.12}$$

$$I(X;Y)=H(Y)-H(Y|X)\,, \tag{16.13}$$

$$I(X;Y)=H(X)+H(Y)-H(X,Y)\,, \tag{16.14}$$

$$I(X;X)=H(X)\,. \tag{16.15}$$

Theorem 16.1.10 (*Section 2.9*): *For a Markov chain:*

1. *Relative entropy $D(\mu_n\|\mu_n')$ decreases with time.*
2. *Relative entropy $D(\mu_n\|\mu)$ between a distribution and the stationary distribution decreases with time.*
3. *Entropy $H(X_n)$ increases if the stationary distribution is uniform.*

4. *The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.*

Theorem 16.1.11 (*Problem 34, Chapter 2*): *Let $X_1, X_2, \dots, X_n$ be i.i.d. $\sim p(x)$. Let $\hat{p}_n$ be the empirical probability mass function of $X_1, X_2, \dots, X_n$. Then*

$$ED(\hat{p}_n \| p) \le ED(\hat{p}_{n-1} \| p) . \tag{16.16}$$

16.2 DIFFERENTIAL ENTROPY

We now review some of the basic properties of differential entropy (Section 9.1).

Definition: The *differential entropy* $h(X_1, X_2, \dots, X_n)$, sometimes written $h(f)$, is defined by

$$h(X_1, X_2, \dots, X_n) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x} . \tag{16.17}$$

The differential entropy for many common densities is given in Table 16.1 (taken from Lazo and Rathie [265]).

Definition: The *relative entropy* between probability densities f and g is

$$D(f \| g) = \int f(\mathbf{x}) \log (f(\mathbf{x})/g(\mathbf{x}))\, d\mathbf{x} . \tag{16.18}$$

The properties of the continuous version of relative entropy are identical to the discrete version. Differential entropy, on the other hand, has some properties that differ from those of discrete entropy. For example, differential entropy may be negative.

We now restate some of the theorems that continue to hold for differential entropy.

Theorem 16.2.1 (*Theorem 9.6.1: Conditioning reduces entropy*): *$h(X|Y) \le h(X)$, with equality iff X and Y are independent.*

Theorem 16.2.2 (*Theorem 9.6.2: Chain rule*):

$$h(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} h(X_i|X_{i-1}, X_{i-2}, \dots, X_1) \le \sum_{i=1}^{n} h(X_i) \tag{16.19}$$

with equality iff $X_1, X_2, \dots, X_n$ are independent.

TABLE 16.1. Table of differential entropies. All entropies are in nats. $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt$. $\psi(z) = \frac{d}{dz}\Gamma(z)$. γ = Euler's constant $= 0.57721566\ldots$. $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$.

Distribution		*Entropy (in nats)*
Name	*Density*	
Beta	$f(w) = \frac{x^{p-1}(1-x)^{q-1}}{B(p, q)}$, $0 \le x \le 1,\ p, q > 0$	$\ln B(p, q) - (p-1)$ $\times [\psi(p) - \psi(p+q)]$ $-(q-1)[\psi(q) - \psi(p+q)]$
Cauchy	$f(x) = \frac{\lambda}{\pi} \frac{1}{\lambda^2 + x^2}$, $-\infty < x < \infty,\ \lambda > 0$	$\ln(4\pi\lambda)$
Chi	$f(x) = \frac{2}{2^{n/2}\sigma^n \Gamma(n/2)}\, x^{n-1} e^{-\frac{x^2}{2\sigma^2}}$, $x > 0,\ n > 0$	$\ln \frac{\sigma\Gamma(n/2)}{\sqrt{2}} - \frac{n-1}{2}\, \psi(\frac{n}{2}) + \frac{n}{2}$
Chi-squared	$f(x) = \frac{1}{2^{n/2}\sigma^n \Gamma(n/2)}\, x^{\frac{n}{2}-1} e^{-\frac{x}{2\sigma^2}}$, $x > 0,\ n > 0$	$\ln 2\sigma^2\Gamma(n/2)$ $-(1 - \frac{n}{2})\psi(\frac{n}{2}) + \frac{n}{2}$
Erlang	$f(x) = \frac{\beta^n}{(n-1)!}\, x^{n-1} e^{-\beta x}$, $x, \beta > 0,\ n > 0$	$(1-n)\psi(n) + \ln \frac{\Gamma(n)}{\beta} + n$
Exponential	$f(x) = \frac{1}{\lambda}\, e^{-\frac{x}{\lambda}}, \quad x, \lambda > 0$	$1 + \ln \lambda$
F	$f(x) = \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}}}{B(\frac{n_1}{2}, \frac{n_2}{2})}\, \frac{x^{\frac{n_1}{2}-1}}{(n_2 + n_1 x)^{\frac{n_1+n_2}{2}}}$,	$\ln \frac{n_1}{n_2}\, B(\frac{n_1}{2}, \frac{n_2}{2})$ $+ (1 - \frac{n_1}{2})\psi(\frac{n_1}{2})$ $- (1 + \frac{n_2}{2})\psi(\frac{n_2}{2})$ $+ \frac{n_1+n_2}{2}\psi(\frac{n_1+n_2}{2})$
Gamma	$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)}$, $x, \alpha, \beta > 0$	$\ln(\beta\Gamma(\alpha)) + (1-\alpha)$ $\times\ \psi(\alpha) + \alpha$
Laplace	$f(x) = \frac{1}{2\lambda}\, e^{-\frac{\|x-\theta\|}{\lambda}}$, $-\infty < x, \theta < \infty,\ \lambda > 0$	$1 + \ln(2\lambda)$

TABLE 16.1. *(Continued)*

Distribution		
Name	*Density*	*Entropy (in nats)*
Logistic	$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$, $-\infty < x < \infty$	2
Lognormal	$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}$, $x > 0, -\infty < m < \infty, \sigma > 0$	$m + \frac{1}{2}\ln(2\pi e \sigma^2)$
Maxwell-Boltzmann	$f(x) = 4\pi^{-\frac{1}{2}} \beta^{\frac{3}{2}} x^2 e^{-\beta x^2}$, $x, \beta > 0$	$\frac{1}{2}\ln\frac{\pi}{\beta} + \gamma - \frac{1}{2}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x, \mu < \infty, \sigma > 0$	$\frac{1}{2}\ln(2\pi e \sigma^2)$
Generalized normal	$f(x) = \frac{2\beta^{\frac{\alpha}{2}}}{\Gamma(\frac{\alpha}{2})} x^{\alpha-1} e^{-\beta x^2}$, $x, \alpha, \beta > 0$	$\ln\frac{\Gamma(\frac{\alpha}{2})}{2\beta^{\frac{1}{2}}} - \frac{\alpha-1}{2}\psi(\frac{\alpha}{2}) + \frac{\alpha}{2}$
Pareto	$f(x) = \frac{ak^a}{x^{a+1}}, \quad x \ge k > 0, a > 0$	$\ln\frac{k}{a} + 1 + \frac{1}{a}$
Rayleigh	$f(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$, $x, b > 0$	$1 + \ln\frac{\beta}{\sqrt{2}} + \frac{\gamma}{2}$
Student-*t*	$f(x) = \frac{(1+x^2/n)^{-\frac{n+1}{2}}}{\sqrt{n}B(\frac{1}{2}, \frac{n}{2})}$, $-\infty < x < \infty, n > 0$	$\frac{n+1}{2}\psi(\frac{n+1}{2}) - \psi(\frac{n}{2})$ $+ \ln\sqrt{n}B(\frac{1}{2}, \frac{n}{2})$
Triangular	$f(x) = \begin{cases} \frac{2x}{a} & 0 \le x \le a \\ \frac{2(1-x)}{1-a} & a \le x \le 1 \end{cases}$	$\frac{1}{2} - \ln 2$
Uniform	$f(x) = \frac{1}{\beta - \alpha}, \quad \alpha \le x \le \beta$	$\ln(\beta - \alpha)$
Weilbull	$f(x) = \frac{c}{\alpha} x^{c-1} e^{-\frac{x^c}{\alpha}}$, $x, c, \alpha > 0$	$\frac{(c-1)\gamma}{c} + \ln\frac{\alpha^{1/c}}{c} + 1$

Lemma 16.2.1: *If X and Y are independent, then $h(X+Y) \geq h(X)$.*

Proof: $h(X+Y) \geq h(X+Y|Y) = h(X|Y) = h(X)$. □

Theorem 16.2.3 (*Theorem 9.6.5*): *Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^t$, i.e., $K_{ij} = EX_iX_j$, $1 \leq i, j \leq n$. Then*

$$h(\mathbf{X}) \leq \frac{1}{2}\log(2\pi e)^n|K|, \tag{16.20}$$

with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.

16.3 BOUNDS ON ENTROPY AND RELATIVE ENTROPY

In this section, we revisit some of the bounds on the entropy function. The most useful is Fano's inequality, which is used to bound away from zero the probability of error of the best decoder for a communication channel at rates above capacity.

Theorem 16.3.1 (*Theorem 2.11.1: Fano's inequality*): *Given two random variables X and Y, let P_e be the probability of error in the best estimator of X given Y. Then*

$$H(P_e) + P_e\log(|\mathcal{X}| - 1) \geq H(X|Y). \tag{16.21}$$

Consequently, if $H(X|Y) > 0$, then $P_e > 0$.

Theorem 16.3.2 (*$\mathcal{L}_1$ bound on entropy*): *Let p and q be two probability mass functions on $\mathcal{X}$ such that*

$$\|p-q\|_1 = \sum_{x\in\mathcal{X}} |p(x) - q(x)| \leq \frac{1}{2}. \tag{16.22}$$

Then

$$|H(p) - H(q)| \leq -\|p-q\|_1 \log\frac{\|p-q\|_1}{|\mathcal{X}|}. \tag{16.23}$$

Proof: Consider the function $f(t) = -t\log t$ shown in Figure 16.1. It can be verified by differentiation that the function $f(\cdot)$ is concave. Also $f(0) = f(1) = 0$. Hence the function is positive between 0 and 1.

Consider the chord of the function from t to $t+\nu$ (where $\nu \leq \frac{1}{2}$). The maximum absolute slope of the chord is at either end (when $t = 0$ or $1-\nu$). Hence for $0 \leq t \leq 1-\nu$, we have

$$|f(t) - f(t+\nu)| \leq \max\{f(\nu), f(1-\nu)\} = -\nu\log\nu. \tag{16.24}$$

Let $r(x) = |p(x) - q(x)|$. Then

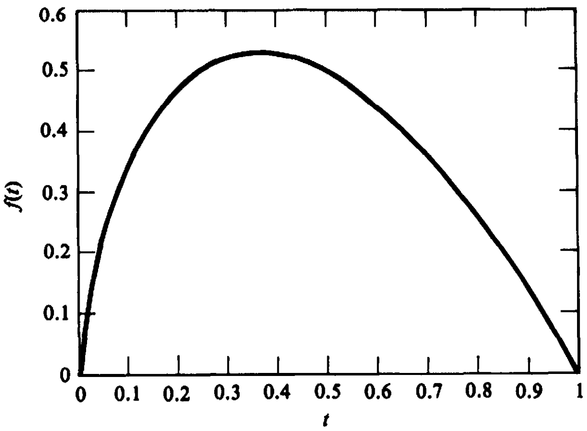

Figure 16.1. The function $f(t) = -t \log t$.

$$|H(p) - H(q)| = \left|\sum_{x\in\mathscr{X}} (-p(x)\log p(x) + q(x)\log q(x))\right| \tag{16.25}$$

$$\le \sum_{x\in\mathscr{X}} |(-p(x)\log p(x) + q(x)\log q(x))| \tag{16.26}$$

$$\le \sum_{x\in\mathscr{X}} -r(x)\log r(x) \tag{16.27}$$

$$= \|p-q\|_1 \sum_{x\in\mathscr{X}} -\frac{r(x)}{\|p-q\|_1} \log \frac{r(x)}{\|p-q\|_1} \|p-q\|_1 \tag{16.28}$$

$$= -\|p-q\|_1 \log \|p-q\|_1 + \|p-q\|_1 H\left(\frac{r(x)}{\|p-q\|_1}\right) \tag{16.29}$$

$$\le -\|p-q\|_1 \log \|p-q\|_1 + \|p-q\|_1 \log|\mathscr{X}|, \tag{16.30}$$

where (16.27) follows from (16.24). □

We can use the concept of differential entropy to obtain a bound on the entropy of a distribution.

Theorem 16.3.3 (*Theorem 9.7.1*):

$$H(p_1, p_2, \dots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). \tag{16.31}$$

Finally, relative entropy is stronger than the $\mathscr{L}_1$ norm in the following sense:

Lemma 16.3.1 (*Lemma 12.6.1*):

$$D(p_1\|p_2) \ge \frac{1}{2\ln 2}\|p_1 - p_2\|_1^2. \tag{16.32}$$

16.4 INEQUALITIES FOR TYPES

The method of types is a powerful tool for proving results in large deviation theory and error exponents. We repeat the basic theorems:

Theorem 16.4.1 (*Theorem 12.1.1*): *The number of types with denominator n is bounded by*

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}. \tag{16.33}$$

Theorem 16.4.2 (*Theorem 12.1.2*): *If $X_1, X_2, \dots, X_n$ are drawn i.i.d. according to $Q(x)$, then the probability of x^n depends only on its type and is given by*

$$Q^n(x^n) = 2^{-n(H(P_{x^n}) + D(P_{x^n} \| Q))}. \tag{16.34}$$

Theorem 16.4.3 (*Theorem 12.1.3: Size of a type class T(P)*): *For any type $P \in \mathcal{P}_n$,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}. \tag{16.35}$$

Theorem 16.4.4 (*Theorem 12.1.4*): *For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P\|Q)}$ to first order in the exponent. More precisely,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \le Q^n(T(P)) \le 2^{-nD(P\|Q)}. \tag{16.36}$$

16.5 ENTROPY RATES OF SUBSETS

We now generalize the chain rule for differential entropy. The chain rule provides a bound on the entropy rate of a collection of random variables in terms of the entropy of each random variable:

$$h(X_1, X_2, \dots, X_n) \le \sum_{i=1}^{n} h(X_i). \tag{16.37}$$

We extend this to show that the entropy per element of a subset of a set of random variables decreases as the size of the set increases. This is not true for each subset but is true on the average over subsets, as expressed in the following theorem.

Definition: Let $(X_1, X_2, \dots, X_n)$ have a density, and for every $S \subseteq \{1, 2, \dots, n\}$, denote by $X(S)$ the subset $\{X_i : i \in S)$. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{h(X(S))}{k} . \tag{16.38}$$

Here $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $\{X_1, X_2, \ldots, X_n\}$.

The following theorem by Han [130] says that the average entropy decreases monotonically in the size of the subset.

Theorem 16.5.1:

$$h_1^{(n)} \ge h_2^{(n)} \ge \cdots \ge h_n^{(n)} . \tag{16.39}$$

Proof: We first prove the last inequality, $h_n^{(n)} \le h_{n-1}^{(n)}$. We write

$$\begin{aligned} h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{n-1}) + h(X_n | X_1, X_2, \ldots, X_{n-1}) , \\ h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{n-2}, X_n) \\ &\quad + h(X_{n-1} | X_1, X_2, \ldots, X_{n-2}, X_n) \\ &\le h(X_1, X_2, \ldots, X_{n-2}, X_n) + h(X_{n-1} | X_1, X_2, \ldots, X_{n-2}) , \\ &\vdots \\ h(X_1, X_2, \ldots, X_n) &\le h(X_2, X_3, \ldots, X_n) + h(X_1) . \end{aligned}$$

Adding these n inequalities and using the chain rule, we obtain

$$nh(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + h(X_1, X_2, \ldots, X_n) \tag{16.40}$$

or

$$\frac{1}{n} h(X_1, X_2, \ldots, X_n) \le \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1} , \tag{16.41}$$

which is the desired result $h_n^{(n)} \le h_{n-1}^{(n)}$.

We now prove that $h_k^{(n)} \le h_{k-1}^{(n)}$ for all $k \le n$ by first conditioning on a k-element subset, and then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $h_k^{(k)} \le h_{k-1}^{(k)}$, and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

Theorem 16.5.2: *Let $r > 0$, and define*

$$t_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} e^{\frac{rh(X(S))}{k}} . \tag{16.42}$$

Then

$$t_1^{(n)} \geq t_2^{(n)} \geq \cdots \geq t_n^{(n)} . \tag{16.43}$$

Proof: Starting from (16.41) in the previous theorem, we multiply both sides by r, exponentiate and then apply the arithmetic mean geometric mean inequality to obtain

$$e^{\frac{1}{n} rh(X_1, X_2, \ldots, X_n)} \leq e^{\frac{1}{n}\sum_{i=1}^{n} \frac{rh(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1}} \tag{16.44}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} e^{\frac{rh(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1}} \quad \text{for all } r \geq 0 , \tag{16.45}$$

which is equivalent to $t_n^{(n)} \leq t_{n-1}^{(n)}$. Now we use the same arguments as in the previous theorem, taking an average over all subsets to prove the result that for all $k \leq n$, $t_k^{(n)} \leq t_{k-1}^{(n)}$. □

Definition: The average *conditional entropy rate per element* for all subsets of size k is the average of the above quantities for k-element subsets of $\{1, 2, \ldots, n\}$, i.e.,

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{h(X(S)|X(S^c))}{k} . \tag{16.46}$$

Here $g_k(S)$ is the entropy per element of the set S conditional on the elements of the set S^c. When the size of the set S increases, one can expect a greater dependence among the elements of the set S, which explains Theorem 16.5.1.

In the case of the conditional entropy per element, as k increases, the size of the conditioning set S^c decreases and the entropy of the set S increases. The increase in entropy per element due to the decrease in conditioning dominates the decrease due to additional dependence among the elements, as can be seen from the following theorem due to Han [130]. Note that the conditional entropy ordering in the following theorem is the reverse of the unconditional entropy ordering in Theorem 16.5.1.

Theorem 16.5.3:

$$g_1^{(n)} \leq g_2^{(n)} \leq \cdots \leq g_n^{(n)} . \tag{16.47}$$

Proof: The proof proceeds on lines very similar to the proof of the theorem for the unconditional entropy per element for a random subset. We first prove that $g_n^{(n)} \geq g_{n-1}^{(n)}$ and then use this to prove the rest of the inequalities.

By the chain rule, the entropy of a collection of random variables is less than the sum of the entropies, i.e.,

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i). \tag{16.48}$$

Subtracting both sides of this inequality from $nh(X_1, X_2, \ldots, X_n)$, we have

$$(n-1)h(X_1, X_2, \ldots, X_n) \geq \sum_{i=1}^{n} (h(X_1, X_2, \ldots, X_n) - h(X_i)) \tag{16.49}$$

$$= \sum_{i=1}^{n} h(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n | X_i). \tag{16.50}$$

Dividing this by $n(n-1)$, we obtain

$$\frac{h(X_1, X_2, \ldots, X_n)}{n} \geq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n | X_i)}{n-1}, \tag{16.51}$$

which is equivalent to $g_n^{(n)} \geq g_{n-1}^{(n)}$.

We now prove that $g_k^{(n)} \geq g_{k-1}^{(n)}$ for all $k \leq n$ by first conditioning on a k-element subset, and then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $g_k^{(k)} \geq g_{k-1}^{(k)}$, and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

Theorem 16.5.4: *Let*

$$f_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} \frac{I(X(S); X(S^c))}{k}. \tag{16.52}$$

Then

$$f_1^{(n)} \geq f_2^{(n)} \geq \cdots \geq f_n^{(n)}. \tag{16.53}$$

Proof: The theorem follows from the identity $I(X(S); X(S^c)) = h(X(S)) - h(X(S)|X(S^c))$ and Theorems 16.5.1 and 16.5.3. □

16.6 ENTROPY AND FISHER INFORMATION

The differential entropy of a random variable is a measure of its descriptive complexity. The Fisher information is a measure of the minimum error in estimating a parameter of a distribution. In this section, we will derive a relationship between these two fundamental quantities and use this to derive the entropy power inequality.

Let X be any random variable with density $f(x)$. We introduce a location parameter θ and write the density in a parametric form as $f(x-\theta)$. The Fisher information (Section 12.11) with respect to θ is given by

$$J(\theta)=\int_{-\infty}^{\infty} f(x-\theta)\left[\frac{\partial}{\partial\theta}\ln f(x-\theta)\right]^2 dx\,. \tag{16.54}$$

In this case, differentiation with respect to x is equivalent to differentiation with respect to θ. So we can write the Fisher information as

$$J(X)=\int_{-\infty}^{\infty} f(x-\theta)\left[\frac{\partial}{\partial x}\ln f(x-\theta)\right]^2 dx=\int_{-\infty}^{\infty} f(x)\left[\frac{\partial}{\partial x}\ln f(x)\right]^2 dx\,, \tag{16.55}$$

which we can rewrite as

$$J(X)=\int_{-\infty}^{\infty} f(x)\left[\frac{\frac{\partial}{\partial x}f(x)}{f(x)}\right]^2 dx\,. \tag{16.56}$$

We will call this the Fisher information of the distribution of X. Notice that, like entropy, it is a function of the density.

The importance of Fisher information is illustrated in the following theorem:

Theorem 16.6.1 (*Theorem 12.11.1: Cramér-Rao inequality*): *The mean squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information, i.e.,*

$$\operatorname{var}(T)\ge\frac{1}{J(\theta)}\,. \tag{16.57}$$

We now prove a fundamental relationship between the differential entropy and the Fisher information:

Theorem 16.6.2 (*de Bruijn's identity: Entropy and Fisher information*): *Let X be any random variable with a finite variance with a* *density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Then*

$$\frac{\partial}{\partial t} h_e(X+\sqrt{t}Z) = \frac{1}{2} J(X+\sqrt{t}Z)\,, \tag{16.58}$$

where h_e is the differential entropy to base e. In particular, if the limit exists as $t \to 0$,

$$\left.\frac{\partial}{\partial t} h_e(X+\sqrt{t}Z)\right|_{t=0} = \frac{1}{2} J(X)\,. \tag{16.59}$$

Proof: Let $Y_t = X + \sqrt{t}Z$. Then the density of Y_t is

$$g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dx\,. \tag{16.60}$$

Then

$$\frac{\partial}{\partial t} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{\partial}{\partial t}\left[\frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}}\right] dx \tag{16.61}$$

$$= \int_{-\infty}^{\infty} f(x)\left[-\frac{1}{2t}\frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} + \frac{(y-x)^2}{2t^2}\frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}}\right] dx\,. \tag{16.62}$$

We also calculate

$$\frac{\partial}{\partial y} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \frac{\partial}{\partial y}\left[e^{-\frac{(y-x)^2}{2t}}\right] dx \tag{16.63}$$

$$= \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}}\left[-\frac{y-x}{t} e^{-\frac{(y-x)^2}{2t}}\right] dx \tag{16.64}$$

and

$$\frac{\partial^2}{\partial y^2} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \frac{\partial}{\partial y}\left[-\frac{y-x}{t} e^{-\frac{(y-x)^2}{2t}}\right] dx \tag{16.65}$$

$$= \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}}\left[-\frac{1}{t} e^{-\frac{(y-x)^2}{2t}} + \frac{(y-x)^2}{t^2} e^{-\frac{(y-x)^2}{2t}}\right] dx. \tag{16.66}$$

Thus

$$\frac{\partial}{\partial t} g_t(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y)\,. \tag{16.67}$$

We will use this relationship to calculate the derivative of the entropy of Y_t, where the entropy is given by

$$h_e(Y_t) = -\int_{-\infty}^{\infty} g_t(y) \ln g_t(y)\, dy\,. \tag{16.68}$$

Differentiating, we obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\int_{-\infty}^{\infty} \frac{\partial}{\partial t} g_t(y)\, dy - \int_{-\infty}^{\infty} \frac{\partial}{\partial t} g_t(y) \ln g_t(y)\, dy \tag{16.69}$$

$$= -\frac{\partial}{\partial t} \int_{-\infty}^{\infty} g_t(y)\, dy - \frac{1}{2} \int_{-\infty}^{\infty} \frac{\partial^2}{\partial y^2} g_t(y) \ln g_t(y)\, dy\,. \tag{16.70}$$

The first term is zero since $\int g_t(y)\, dy = 1$. The second term can be integrated by parts to obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\frac{1}{2}\left[\frac{\partial g_t(y)}{\partial y} \ln g_t(y)\right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial y} g_t(y)\right]^2 \frac{1}{g_t(y)}\, dy\,. \tag{16.71}$$

The second term in (16.71) is $\frac{1}{2}J(Y_t)$. So the proof will be complete if we show that the first term in (16.71) is zero. We can rewrite the first term as

$$\frac{\partial g_t(y)}{\partial y} \ln g_t(y) = \left[\frac{\frac{\partial g_t(y)}{\partial y}}{\sqrt{g_t(y)}}\right] [2\sqrt{g_t(y)} \ln \sqrt{g_t(y)}]\,. \tag{16.72}$$

The square of the first factor integrates to the Fisher information, and hence must be bounded as $y \to \pm\infty$. The second factor goes to zero since $x \ln x \to 0$ as $x \to 0$ and $g_t(y) \to 0$ as $y \to \pm\infty$. Hence the first term in (16.71) goes to 0 at both limits and the theorem is proved.

In the proof, we have exchanged integration and differentiation in (16.61), (16.63), (16.65) and (16.69). Strict justification of these exchanges requires the application of the bounded convergence and mean value theorems; the details can be found in Barron [15]. □

This theorem can be used to prove the entropy power inequality, which gives a lower bound on the entropy of a sum of independent random variables.

Theorem 16.6.3: (*Entropy power inequality*): *If* **X** *and* **Y** *are independent random n-vectors with densities, then*

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}\,. \tag{16.73}$$

We outline the basic steps in the proof due to Stam [257] and Blachman [34]. The next section contains a different proof.

Stam's proof of the entropy power inequality is based on a perturbation argument. Let $X_t = X + \sqrt{f(t)}Z_1$, $Y_t = Y + \sqrt{g(t)}Z_2$, where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$ random variables. Then the entropy power inequality reduces to showing that $s(0) \le 1$, where we define

$$s(t) = \frac{2^{2h(X_t)} + 2^{2h(Y_t)}}{2^{2h(X_t+Y_t)}} . \qquad (16.74)$$

If $f(t) \to \infty$ and $g(t) \to \infty$ as $t \to \infty$, then it is easy to show that $s(\infty) = 1$. If, in addition, $s'(t) \ge 0$ for $t \ge 0$, this implies that $s(0) \le 1$. The proof of the fact that $s'(t) \ge 0$ involves a clever choice of the functions $f(t)$ and $g(t)$, an application of Theorem 16.6.2 and the use of a convolution inequality for Fisher information,

$$\frac{1}{J(X+Y)} \ge \frac{1}{J(X)} + \frac{1}{J(Y)} . \qquad (16.75)$$

The entropy power inequality can be extended to the vector case by induction. The details can be found in papers by Stam [257] and Blachman [34].

16.7 THE ENTROPY POWER INEQUALITY AND THE BRUNN-MINKOWSKI INEQUALITY

The entropy power inequality provides a lower bound on the differential entropy of a sum of two independent random vectors in terms of their individual differential entropies. In this section, we restate and outline a new proof of the entropy power inequality. We also show how the entropy power inequality and the Brunn-Minkowski inequality are related by means of a common proof.

We can rewrite the entropy power inequality in a form that emphasizes its relationship to the normal distribution. Let X and Y be two independent random variables with densities, and let X' and Y' be independent normals with the same entropy as X and Y, respectively. Then $2^{2h(X)} = 2^{2h(X')} = (2\pi e)\sigma^2_{X'}$ and similarly $2^{2h(Y)} = (2\pi e)\sigma^2_{Y'}$. Hence the entropy power inequality can be rewritten as

$$2^{2h(X+Y)} \ge (2\pi e)(\sigma^2_{X'} + \sigma^2_{Y'}) = 2^{2h(X'+Y')} , \qquad (16.76)$$

since X' and Y' are independent. Thus we have a new statement of the entropy power inequality:

Theorem 16.7.1 (*Restatement of the entropy power inequality*): *For two independent random variables X and Y,*

$$h(X+Y) \geq h(X'+Y'), \tag{16.77}$$

where X' and Y' are independent normal random variables with $h(X') = h(X)$ and $h(Y') = h(Y)$.

This form of the entropy power inequality bears a striking resemblance to the Brunn-Minkowski inequality, which bounds the volume of set sums.

Definition: The *set sum* $A+B$ of two sets $A, B \subset \mathcal{R}^n$ is defined as the set $\{x+y : x \in A, y \in B\}$.

Example 16.7.1: The set sum of two spheres of radius 1 at the origin is a sphere of radius 2 at the origin.

Theorem 16.7.2 (*Brunn-Minkowski inequality*): *The volume of the set sum of two sets A and B is greater than the volume of the set sum of two spheres A' and B' with the same volumes as A and B, respectively, i.e.,*

$$V(A+B) \overset{C}{\geq} V(A'+B'), \tag{16.78}$$

where A' and B' are spheres with $V(A') = V(A)$ and $V(B') = V(B)$.

The similarity between the two theorems was pointed out in [58]. A common proof was found by Dembo [87] and Lieb, starting from a strengthened version of Young's inequality. The same proof can be used to prove a range of inequalities which includes the entropy power inequality and the Brunn-Minkowski inequality as special cases. We will begin with a few definitions.

Definition: Let f and g be two densities over $\mathcal{R}^n$ and let $f*g$ denote the convolution of the two densities. Let the $\mathcal{L}_r$ norm of the density be defined by

$$\|f\|_r = \left(\int f^r(x)\,dx\right)^{1/r}. \tag{16.79}$$

Lemma 16.7.1 (*Strengthened Young's inequality*): *For any two densities f and g,*

$$\|f*g\|_r \leq \left(\frac{C_p C_q}{C_r}\right)^{n/2} \|f\|_p \|g\|_q , \tag{16.80}$$

where
$$\frac{1}{r}=\frac{1}{p}+\frac{1}{q}-1 \tag{16.81}$$

and

$$C_p=\frac{p^{\frac{1}{p}}}{p'^{\frac{1}{p'}}}, \qquad \frac{1}{p}+\frac{1}{p'}=1\,. \tag{16.82}$$

Proof: The proof of this inequality is rather involved; it can be found in [19] and [43]. □

We define a generalization of the entropy:

Definition: The *Renyi entropy* $h_r(X)$ of order r is defined as

$$h_r(X)=\frac{1}{1-r}\log\left[\int f^r(x)\,dx\right] \tag{16.83}$$

for $0<r<\infty$, $r\neq 1$. If we take the limit as $r\rightarrow 1$, we obtain the Shannon entropy function

$$h(X)=h_1(X)=-\int f(x)\log f(x)\,dx\,. \tag{16.84}$$

If we take the limit as $r\rightarrow 0$, we obtain the logarithm of the volume of the support set,

$$h_0(X)=\log(\mu\{x: f(x)>0\})\,. \tag{16.85}$$

Thus the zeroth order Renyi entropy gives the measure of the support set of the density f. We now define the equivalent of the entropy power for Renyi entropies.

Definition: The *Renyi entropy power* $V_r(X)$ of order r is defined as

$$V_r(X)=\begin{cases}[\int f^r(x)\,dx]^{-\frac{2}{n}\frac{r'}{r}}, & 0<r\leq\infty\,, r\neq 1\,, \frac{1}{r}+\frac{1}{r'}=1\\ \exp[\frac{2}{n}h(X)]\,, & r=1\\ \mu(\{x: f(x)>0\})^{\frac{2}{n}}, & r=0\end{cases} \tag{16.86}$$

Theorem 16.7.3: *For two independent random variables X and Y and any $0\leq r<\infty$ and any $0\leq\lambda\leq 1$, we have*

$$\log V_r(X+Y) \geq \lambda \log V_p(X) + (1-\lambda)\log V_q(Y) + H(\lambda)$$
$$+ \left(\frac{1+r}{1-r}\right)\left[H\left(\frac{r+\lambda(1-r)}{1+r}\right) - H\left(\frac{r}{1+r}\right)\right], \qquad (16.87)$$

where $p = \frac{r}{r+\lambda(1-r)}$, $q = \frac{r}{r+(1-\lambda)(1-r)}$ *and* $H(\lambda) = -\lambda\log\lambda - (1-\lambda)\log(1-\lambda)$.

Proof: If we take the logarithm of Young's inequality (16.80), we obtain

$$\frac{1}{r'}\log V_r(X+Y) \geq \frac{1}{p'}\log V_p(X) + \frac{1}{q'}\log V_q(Y)$$
$$+ \log C_r - \log C_p - \log C_q\,. \qquad (16.88)$$

Setting $\lambda = r'/p'$ and using (16.81), we have $1-\lambda = r'/q'$, $p = \frac{r}{r+\lambda(1-r)}$ and $q = \frac{r}{r+(1-\lambda)(1-r)}$. Thus (16.88) becomes

$$\log V_r(X+Y) \geq \lambda\log V_p(X) + (1-\lambda)\log V_q(Y) + \frac{r'}{r}\log r - \log r'$$
$$- \frac{r'}{p}\log p + \frac{r'}{p'}\log p' - \frac{r'}{q}\log q + \frac{r'}{q'}\log q' \qquad (16.89)$$
$$= \lambda\log V_p(X) + (1-\lambda)\log V_q(Y) + \frac{r'}{r}\log r - (\lambda + 1 - \lambda)\log r'$$
$$- \frac{r'}{p}\log p + \lambda\log p' - \frac{r'}{q}\log q + (1-\lambda)\log q' \qquad (16.90)$$
$$= \lambda\log V_p(X) + (1-\lambda)\log V_q(Y) + \frac{1}{r-1}\log r + H(\lambda)$$
$$- \frac{r+\lambda(1-r)}{r-1}\log\frac{r}{r+\lambda(1-r)} \qquad (16.91)$$
$$- \frac{r+(1-\lambda)(1-r)}{r-1}\log\frac{r}{r+(1-\lambda)(1-r)} \qquad (16.92)$$
$$= \lambda\log V_p(X) + (1-\lambda)\log V_q(Y) + H(\lambda)$$
$$+ \left(\frac{1+r}{1-r}\right)\left[H\left(\frac{r+\lambda(1-r)}{1+r}\right) - H\left(\frac{r}{1+r}\right)\right], \qquad (16.93)$$

where the details of the algebra for the last step are omitted. □

The Brunn-Minkowski inequality and the entropy power inequality can then be obtained as special cases of this theorem.

- *The entropy power inequality.* Taking the limit of (16.87) as $r \rightarrow 1$ and setting

$$\lambda = \frac{V_1(X)}{V_1(X) + V_1(Y)}, \tag{16.94}$$

we obtain

$$V_1(X+Y) \geq V_1(X) + V_1(Y), \tag{16.95}$$

which is the entropy power inequality.

- *The Brunn-Minkowski inequality.* Similarly letting $r \rightarrow 0$ and choosing

$$\lambda = \frac{\sqrt{V_0(X)}}{\sqrt{V_0(X)} + \sqrt{V_0(Y)}}, \tag{16.96}$$

we obtain

$$\sqrt{V_0(X+Y)} \geq \sqrt{V_0(X)} + \sqrt{V_0(Y)}, \tag{16.97}$$

Now let A be the support set of X and B be the support set of Y. Then $A + B$ is the support set of $X + Y$, and (16.97) reduces to

$$[\mu(A+B)]^{1/n} \geq [\mu(A)]^{1/n} + [\mu(B)]^{1/n}, \tag{16.98}$$

which is the Brunn-Minkowski inequality.

The general theorem unifies the entropy power inequality and the Brunn-Minkowski inequality, and also introduces a continuum of new inequalities that lie between the entropy power inequality and the Brunn-Minkowski inequality. This furthers strengthens the analogy between entropy power and volume.

16.8 INEQUALITIES FOR DETERMINANTS

Throughout the remainder of this chapter, we will assume that K is a non-negative definite symmetric $n \times n$ matrix. Let $|K|$ denote the determinant of K.

We first prove a result due to Ky Fan [103].

Theorem 16.8.1: $\log|K|$ *is concave.*

Proof: Let X_1 and X_2 be normally distributed n-vectors, $\mathbf{X}_i \sim \mathcal{N}(0, K_i)$, $i = 1, 2$. Let the random variable θ have the distribution

$$\Pr\{\theta = 1\} = \lambda\,, \tag{16.99}$$

$$\Pr\{\theta = 2\} = 1 - \lambda\,, \tag{16.100}$$

for some $0 \le \lambda \le 1$. Let θ, $\mathbf{X}_1$ and $\mathbf{X}_2$ be independent and let $\mathbf{Z} = \mathbf{X}_\theta$. Then $\mathbf{Z}$ has covariance $K_Z = \lambda K_1 + (1-\lambda)K_2$. However, $\mathbf{Z}$ will not be multivariate normal. By first using Theorem 16.2.3, followed by Theorem 16.2.1, we have

$$\begin{aligned} \frac{1}{2}\log(2\pi e)^n |\lambda K_1 + (1-\lambda)K_2| &\ge h(\mathbf{Z}) && (16.101)\\ &\ge h(\mathbf{Z}|\theta) && (16.102)\\ &= \lambda \tfrac{1}{2}\log(2\pi e)^n |K_1| + (1-\lambda)\tfrac{1}{2}\log(2\pi e)^n |K_2|\,. \end{aligned}$$

$$|\lambda K_1 + (1-\lambda)K_2| \ge |K_1|^\lambda |K_2|^{1-\lambda}\,, \tag{16.103}$$

as desired. □

We now give Hadamard's inequality using an information theoretic proof [68].

Theorem 16.8.2 (*Hadamard*): $|K| \le \Pi K_{ii}$, *with equality iff* $K_{ij} = 0$, $i \ne j$.

Proof: Let $\mathbf{X} \sim \mathcal{N}(0, K)$. Then

$$\frac{1}{2}\log(2\pi e)^n |K| = h(X_1, X_2, \ldots, X_n) \le \sum h(X_i) = \sum_{i=1}^{n} \frac{1}{2}\log 2\pi e|K_{ii}|\,, \tag{16.104}$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent, i.e., $K_{ij} = 0$, $i \ne j$. □

We now prove a generalization of Hadamard's inequality due to Szasz [196]. Let $K(i_1, i_2, \ldots, i_k)$ be the $k \times k$ principal submatrix of K formed by the rows and columns with indices $i_1, i_2, \ldots, i_k$.

Theorem 16.8.3 (*Szasz*): *If K is a positive definite $n \times n$ matrix and P_k denotes the product of the determinants of all the principal k-rowed minors of K, i.e.,*

$$P_k = \prod_{1 \le i_1 < i_2 < \cdots < i_k \le n} |K(i_1, i_2, \ldots, i_k)| , \tag{16.105}$$

then

$$P_1 \ge P_2^{1/\binom{n-1}{1}} \ge P_3^{1/\binom{n-1}{2}} \ge \cdots \ge P_n . \tag{16.106}$$

Proof: Let $\mathbf{X} \sim \mathcal{N}(0, K)$. Then the theorem follows directly from Theorem 16.5.1, with the identification $h_k^{(n)} = \frac{1}{2n\binom{n-1}{k-1}} \log P_k + \frac{1}{2} \log 2\pi e$. □

We can also prove a related theorem.

Theorem 16.8.4: *Let K be a positive definite $n \times n$ matrix and let*

$$S_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} |K(i_1, i_2, \ldots, i_k)|^{1/k} . \tag{16.107}$$

Then

$$\frac{1}{n} \operatorname{tr}(K) = S_1^{(n)} \ge S_2^{(n)} \ge \cdots \ge S_n^{(n)} = |K|^{1/n} . \tag{16.108}$$

Proof: This follows directly from the corollary to Theorem 16.5.1, with the identification $t_k^{(n)} = (2\pi e) S_k^{(n)}$ and $r = 2$. □

Theorem 16.8.5: *Let*

$$Q_k = \left(\prod_{S : |S| = k} \frac{|K|}{|K(S^c)|} \right)^{1/k\binom{n}{k}} . \tag{16.109}$$

Then

$$\left(\prod_{i=1}^{n} \sigma_i^2 \right)^{1/n} = Q_1 \le Q_2 \le \cdots \le Q_{n-1} \le Q_n = |K|^{1/n} . \tag{16.110}$$

Proof: The theorem follows immediately from Theorem 16.5.3 and the identification

$$h(X(S)|X(S^c)) = \frac{1}{2} \log(2\pi e)^k \frac{|K|}{|K(S^c)|} . \quad \square \tag{16.111}$$

The outermost inequality, $Q_1 \le Q_n$, can be rewritten as

$$|K| \ge \prod_{i=1}^{n} \sigma_i^2 , \tag{16.112}$$

where $$\sigma_i^2 = \frac{|K|}{|K(1,2\dots,i-1,i+1,\dots,n)|} \tag{16.113}$$

is the minimum mean squared error in the linear prediction of X_i from the remaining X's. Thus σ_i^2 is the conditional variance of X_i given the remaining X_j's if $X_1, X_2, \dots, X_n$ are jointly normal. Combining this with Hadamard's inequality gives upper and lower bounds on the determinant of a positive definite matrix:

Corollary:

$$\prod_i K_{ii} \ge |K| \ge \prod_i \sigma_i^2 . \tag{16.114}$$

Hence the determinant of a covariance matrix lies between the product of the unconditional variances K_{ii} of the random variables X_i and the product of the conditional variances σ_i^2.

We now prove a property of Toeplitz matrices, which are important as the covariance matrices of stationary random processes. A Toeplitz matrix K is characterized by the property that $K_{ij} = K_{rs}$ if $|i-j| = |r-s|$. Let K_k denote the principal minor $K(1, 2, \dots, k)$. For such a matrix, the following property can be proved easily from the properties of the entropy function.

Theorem 16.8.6: *If the positive definite $n \times n$ matrix K is Toeplitz, then*

$$|K_1| \ge |K_2|^{1/2} \ge 0 \cdots \ge |K_{n-1}|^{1/(n-1)} \ge |K_n|^{1/n} \tag{16.115}$$

and $|K_k|/|K_{k-1}|$ is decreasing in k, and

$$\lim_{n\to\infty} |K_n|^{1/n} = \lim_{n\to\infty} \frac{|K_n|}{|K_{n-1}|} . \tag{16.116}$$

Proof: Let $(X_1, X_2, \dots, X_n) \sim \mathcal{N}(0, K_n)$. We observe that

$$h(X_k|X_{k-1}, \dots, X_1) = h(X^k) - h(X^{k-1}) \tag{16.117}$$

$$= \frac{1}{2} \log(2\pi e) \frac{|K_k|}{|K_{k-1}|} . \tag{16.118}$$

Thus the monotonicity of $|K_k|/|K_{k-1}|$ follows from the monotonocity of $h(X_k|X_{k-1}, \dots, X_1)$, which follows from

$$h(X_k|X_{k-1}, \dots, X_1) = h(X_{k+1}|X_k, \dots, X_2) \tag{16.119}$$

$$\ge h(X_{k+1}|X_k, \dots, X_2, X_1) , \tag{16.120}$$

where the equality follows from the Toeplitz assumption and the inequality from the fact that conditioning reduces entropy. Since $h(X_k|X_{k-1}, \dots, X_1)$ is decreasing, it follows that the running averages

$$\frac{1}{k} h(X_1, \dots, X_k) = \frac{1}{k} \sum_{i=1}^{k} h(X_i|X_{i-1}, \dots, X_1) \tag{16.121}$$

are decreasing in k. Then (16.115) follows from $h(X_1, X_2, \dots, X_k) = \frac{1}{2}\log(2\pi e)^k |K_k|$. □

Finally, since $h(X_n|X_{n-1}, \dots, X_1)$ is a decreasing sequence, it has a limit. Hence by the Cesáro mean theorem,

$$\lim_{n\to\infty} \frac{h(X_1, X_2, \dots, X_n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h(X_k|X_{k-1}, \dots, X_1)$$

$$= \lim_{n\to\infty} h(X_n|X_{n-1}, \dots, X_1). \tag{16.122}$$

Translating this to determinants, one obtains

$$\lim_{n\to\infty} |K_n|^{1/n} = \lim_{n\to\infty} \frac{|K_n|}{|K_{n-1}|} . \tag{16.123}$$

Theorem 16.8.7 (*Minkowski inequality [195]*):

$$|K_1 + K_2|^{1/n} \ge |K_1|^{1/n} + |K_2|^{1/n} . \tag{16.124}$$

Proof: Let $\mathbf{X}_1, \mathbf{X}_2$ be independent with $\mathbf{X}_i \sim \mathcal{N}(0, K_i)$. Noting that $\mathbf{X}_1 + \mathbf{X}_2 \sim \mathcal{N}(0, K_1 + K_2)$ and using the entropy power inequality (Theorem 16.6.3) yields

$$(2\pi e)|K_1 + K_2|^{1/n} = 2^{(2/n)h(\mathbf{X}_1 + \mathbf{X}_2)} \tag{16.125}$$

$$\ge 2^{(2/n)h(\mathbf{X}_1)} + 2^{(2/n)h(\mathbf{X}_2)} \tag{16.126}$$

$$= (2\pi e)|K_1|^{1/n} + (2\pi e)|K_2|^{1/n}. \quad \square \tag{16.127}$$

16.9 INEQUALITIES FOR RATIOS OF DETERMINANTS

We now prove similar inequalities for ratios of determinants. Before developing the next theorem, we make an observation about minimum mean squared error linear prediction. If $(X_1, X_2, \dots, X_n) \sim \mathcal{N}(0, K_n)$, we know that the conditional density of X_n given $(X_1, X_2, \dots, X_{n-1})$ is univariate normal with mean linear in $X_1, X_2, \ldots, X_{n-1}$ and conditional variance σ_n^2. Here σ_n^2 is the minimum mean squared error $E(X_n - \hat{X}_n)^2$ over all linear estimators $\hat{X}_n$ based on $X_1, X_2, \ldots, X_{n-1}$.

Lemma 16.9.1: $\sigma_n^2 = |K_n|/|K_{n-1}|$.

Proof: Using the conditional normality of X_n, we have

$$\frac{1}{2}\log 2\pi e\sigma_n^2 = h(X_n|X_1, X_2, \ldots, X_{n-1}) \qquad (16.128)$$

$$= h(X_1, X_2, \ldots, X_n) - h(X_1, X_2, \ldots, X_{n-1}) \qquad (16.129)$$

$$= \frac{1}{2}\log(2\pi e)^n|K_n| - \frac{1}{2}\log(2\pi e)^{n-1}\,|K_{n-1}| \qquad (16.130)$$

$$= \frac{1}{2}\log 2\pi e|K_n|/|K_{n-1}|\,. \quad \square \qquad (16.131)$$

Minimization of σ_n^2 over a set of allowed covariance matrices $\{K_n\}$ is aided by the following theorem. Such problems arise in maximum entropy spectral density estimation.

Theorem 16.9.1 *(Bergstrøm [23])*: $\log(|K_n|/|K_{n-p}|)$ *is concave in* K_n.

Proof: We remark that Theorem 16.8.1 cannot be used because $\log(|K_n|/|K_{n-p}|)$ is the difference of two concave functions. Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim \mathcal{N}(0, S_n)$, $\mathbf{X}_2 \sim \mathcal{N}(0, T_n)$, $\Pr\{\theta = 1\} = \lambda = 1 - \Pr\{\theta = 2\}$ and let $\mathbf{X}_1, \mathbf{X}_2, \theta$ be independent. The covariance matrix K_n of $\mathbf{Z}$ is given by

$$K_n = \lambda S_n + (1-\lambda)T_n\,. \qquad (16.132)$$

The following chain of inequalities proves the theorem:

$$\lambda\,\frac{1}{2}\log(2\pi e)^p|S_n|/|S_{n-p}| + (1-\lambda)\,\frac{1}{2}\log(2\pi e)^p|T_n|/|T_{n-p}|$$

$$\overset{(a)}{=} \lambda h(X_{1,n}, X_{1,n-1}, \ldots, X_{1,n-p+1}|X_{1,1}, \ldots, X_{1,n-p})$$

$$+ (1-\lambda)h(X_{2,n}, X_{2,n-1}, \ldots, X_{2,n-p+1}|X_{2,1}, \ldots, X_{2,n-p}) \qquad (16.133)$$

$$= h(Z_n, Z_{n-1}, \ldots, Z_{n-p+1}|Z_1, \ldots, Z_{n-p}, \theta) \qquad (16.134)$$

$$\overset{(b)}{\leq} h(Z_n, Z_{n-1}, \ldots, Z_{n-p+1}|Z_1, \ldots, Z_{n-p}) \qquad (16.135)$$

$$\overset{(c)}{\leq} \frac{1}{2}\log(2\pi e)^p\,\frac{|K_n|}{|K_{n-p}|}\,, \qquad (16.136)$$

where(a) follows from $h(X_n, X_{n-1}, \ldots, X_{n-p+1}|X_1, \ldots, X_{n-p}) = h(X_1, \ldots, X_n) - h(X_1, \ldots, X_{n-p})$, (b) from the conditioning lemma, and (c) follows from a conditional version of Theorem 16.2.3. □

Theorem 16.9.2 (*Bergstrøm [23]*): *$|K_n|/|K_{n-1}|$ is concave in K_n.*

Proof: Again we use the properties of Gaussian random variables. Let us assume that we have two independent Gaussian random n-vectors, $\mathbf{X} \sim \mathcal{N}(0, A_n)$ and $\mathbf{Y} \sim \mathcal{N}(0, B_n)$. Let $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$.

Then

$$\frac{1}{2} \log 2\pi e \frac{|A_n + B_n|}{|A_{n-1} + B_{n-1}|} \overset{(a)}{=} h(Z_n|Z_{n-1}, Z_{n-2}, \ldots, Z_1) \tag{16.137}$$

$$\overset{(b)}{\geq} h(Z_n|Z_{n-1}, Z_{n-2}, \ldots, Z_1, X_{n-1}, X_{n-2}, \ldots, X_1, Y_{n-1}, Y_{n-2}, \ldots, Y_1) \tag{16.138}$$

$$\overset{(c)}{=} h(X_n + Y_n|X_{n-1}, X_{n-2}, \ldots, X_1, Y_{n-1}, Y_{n-2}, \ldots, Y_1) \tag{16.139}$$

$$\overset{(d)}{=} E \frac{1}{2} \log[2\pi e \operatorname{Var}(X_n + Y_n|X_{n-1}, X_{n-2}, \ldots, X_1, Y_{n-1}, Y_{n-2}, \ldots, Y_1)] \tag{16.140}$$

$$\overset{(e)}{=} E \frac{1}{2} \log[2\pi e(\operatorname{Var}(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) + \operatorname{Var}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1))] \tag{16.141}$$

$$\overset{(f)}{=} E \frac{1}{2} \log\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|} + \frac{|B_n|}{|B_{n-1}|}\right)\right) \tag{16.142}$$

$$= \frac{1}{2} \log\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|} + \frac{|B_n|}{|B_{n-1}|}\right)\right), \tag{16.143}$$

where

(a) follows from Lemma 16.9.1,
(b) from the fact the conditioning decreases entropy,
(c) from the fact that Z is a function of X and Y,
(d) since $X_n + Y_n$ is Gaussian conditioned on $X_1, X_2, \ldots, X_{n-1}, Y_1, Y_2, \ldots, Y_{n-1}$, and hence we can express its entropy in terms of its variance,
(e) from the independence of X_n and Y_n conditioned on the past $X_1, X_2, \ldots, X_{n-1}, Y_1, Y_2, \ldots, Y_{n-1}$, and

(f) follows from the fact that for a set of jointly Gaussian random variables, the conditional variance is constant, independent of the conditioning variables (Lemma 16.9.1).

Setting $A = \lambda S$ and $B = \bar{\lambda} T$, we obtain

$$\frac{|\lambda S_n + \bar{\lambda} T_n|}{|\lambda S_{n-1} + \bar{\lambda} T_{n-1}|} \geq \lambda \frac{|S_n|}{|S_{n-1}|} + \bar{\lambda} \frac{|T_n|}{|T_{n-1}|}, \tag{16.144}$$

i.e., $|K_n|/|K_{n-1}|$ is concave. Simple examples show that $|K_n|/|K_{n-p}|$ is not necessarily concave for $p \geq 2$. □

A number of other determinant inequalities can be proved by these techniques. A few of them are found in the exercises.

OVERALL SUMMARY

Entropy: $H(X) = -\Sigma\, p(x) \log p(x)$.

Relative entropy: $D(p\|q) = \Sigma\, p(x) \log \frac{p(x)}{q(x)}$.

Mutual information: $I(X; Y) = \Sigma\, p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$.

Information inequality: $D(p\|q) \geq 0$.

Asymptotic equipartition property: $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \rightarrow H(X)$.

Data compression: $H(X) \leq L^* < H(X) + 1$.

Kolmogorov complexity: $K(x) = \min_{\mathcal{U}(p)=x} l(p)$.

Channel capacity: $C = \max_{p(x)} I(X; Y)$.

Data transmission:

- $R < C$: Asymptotically error-free communication possible
- $R > C$: Asymptotically error-free communication not possible

Capacity of a white Gaussian noise channel: $C = \frac{1}{2} \log(1 + \frac{P}{N})$.

Rate distortion: $R(D) = \min I(X; \hat{X})$
over all $p(\hat{x}|x)$ such that $E_{p(x)p(\hat{x}|x)} d(X, \hat{X}) \leq D$.

Doubling rate for stock market: $W^* = \max_{\mathbf{b}^*} E \log \mathbf{b}^t \mathbf{X}$.

PROBLEMS FOR CHAPTER 16

1. *Sum of positive definite matrices.* For any two positive definite matrices, K_1 and K_2, show that $|K_1 + K_2| \geq |K_1|$.

2. *Ky Fan inequality [104] for ratios of determinants.* For all $1 \leq p \leq n$, for a positive definite K, show that

$$\frac{|K|}{|K(p+1, p+2, \ldots, n)|} \leq \prod_{i=1}^{p} \frac{|K(i, p+1, p+2, \ldots, n)|}{|K(p+1, p+2, \ldots, n)|} . \qquad (16.145)$$

HISTORICAL NOTES

The entropy power inequality was stated by Shannon [238]; the first formal proofs are due to Stam [257] and Blachman [34]. The unified proof of the entropy power and Brunn-Minkowski inequalities is in Dembo [87].

Most of the matrix inequalities in this chapter were derived using information theoretic methods by Cover and Thomas [59]. Some of the subset inequalities for entropy rates can be found in Han [130].