ON THE CONVEX DUALITY BETWEEN MAXIMUM GROWTH RATES AND MINIMUM INFORMATION RATES

BY
PAUL ALGOET
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 95
AUGUST 1996

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-9628193

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE CONVEX DUALITY BETWEEN MAXIMUM GROWTH RATES AND MINIMUM INFORMATION RATES

Paul Algoet
paul@isl.stanford.edu

Abstract

Let $X$ be a random variable with distribution $P$ on a measurable space $\mathcal{X}$. We investigate the duality between the maximum growth exponent $W_\mathcal{Q}(P) = \sup_{q \in \mathcal{Q}} E_P\{\log q(X)\}$ over a convex family $\mathcal{Q}$ of measurable functions $q(x) \geq 0$ and the minimum information $I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} E_P\{\log(dP/dm(X))\}$ relative to a convex family $\mathcal{M}$ of measures $m(dx)$. We prove that the minimax property $W_\mathcal{Q}(P) = I_\mathcal{M}(P)$ holds if $\mathcal{M}$ is the polar of $\mathcal{Q}$ or vice versa. This is related to results of Calderón and Lozanovskii on interpolation of Banach function spaces. The mutual information $I(X;Y)$ is decomposed into the sum $W_\mathcal{Q}(X;Y) + W_\mathcal{M}(X;Y)$ of two complementary terms that represent the advantage in growth exponent and the advantage in information that accrues from knowledge of the side information $Y$ when making selections in the family $\mathcal{Q}$ and the polar family $\mathcal{M}$. We also discuss existence and uniqueness of log-optimum selections in the families $\mathcal{Q}$ and $\mathcal{M}$, and an alternating minimization algorithm for computing log-optimum selections. This alternating minimization algorithm generalizes Blahut's iterative procedure for computing points on the rate-distortion function of a memoryless source, Cover's algorithm for computing log-optimum portfolios, and the EM algorithm for computing maximum likelihood estimates of mixture distributions. Finally, we show that maximizing determinants or volumes and optimizing spectra are log-optimum selection problems in disguise.

Key words: log-optimum selections in convex families, maximum growth exponent, minimum information, convex duality, minimax theorem, convex corners, Banach function spaces, interpolation theory, Calderón-Lozanovskii construction, value of side information, alternating minimization, rate-distortion theory, gambling.

---

1Parts of this paper were presented at the International Symposia on Information Theory in Kobe, Japan, 1988 and in San Diego, California, 1990.
I. INTRODUCTION

Let $X$ be a random variable with distribution $P$ on a measurable space $\mathcal{X}$. The growth exponent of a nonnegative measurable function $q(x) \geq 0$ is defined as the expectation $E_P\{\log q(X)\}$, provided the expectation exists. The growth exponent will be denoted by $W_q(P)$ or informally by $W_q(X)$:

$$W_q(P) = E_P\{\log q(X)\} = W_q(X).$$

The term “growth exponent” is justified by the strong law of large numbers for products. Indeed, if $\{X_t\}$ is a sequence of independent realizations of $X$ then the product $\prod_{0 \leq t < n} q(X_t)$ grows exponentially fast almost surely with constant limiting rate $W_q(P)$:

$$\prod_{0 \leq t < n} q(X_t) = \exp \left( n \frac{1}{n} \sum_{0 \leq t < n} \log q(X_t) \right) = \exp[n W_q(P) + o(n)],$$

where $o(n) \to 0$ almost surely. The growth exponent may be negative, in which case it describes exponential decay. The compounded product may grow or decrease subexponentially fast if $W_q(P) = 0$, and will grow or decrease superexponentially fast if $W_q(P) = \pm \infty$. If $E_P\{\log q(X)\}$ is not well defined then by convention we set $W_q(P) = -\infty$, so that always

$$W_q(P) = \lim \inf_n \frac{1}{n} \log \left( \prod_{0 \leq t < n} q(X_t) \right).$$

The number $W_q(X)$ is the Lyapunov exponent for the product of the random $1 \times 1$ matrices $q(X_t)$, and perhaps may be called the “characteristic exponent” of $q(X)$. However, in the multivariate case of random $d \times d$ matrices there are several characteristic exponents and we consider mainly their sum, which represents the exponential growth rate of a volume.

We are interested in the maximum growth exponent of $q(X)$ as $q$ ranges over a convex family $Q$ of functions on $\mathcal{X}$. The maximum growth exponent will be denoted by

$$W_Q(P) = \sup_{q \in Q} E_P\{\log q(X)\} = W_Q(X).$$

Any $q^* \in Q$ which attains the maximum is said to be log-optimum in $Q$ (under $P$). Bell and Cover [12] proved that the log-optimum $q^*(x)$ is characterized by Karush-Kuhn-Tucker conditions. The log-optimum $q^*(x)$ need not be unique, but the random variable $q^*(X)$ is unique up to almost sure equivalence. Kieffer [72] proved a result that implies the existence of a log-optimum $q^*(x)$ in some closure of the family $Q$, assuming $W_Q(P)$ is finite.

Maximizing exponential growth rates is the ultimate goal of the information theoretic approach to gambling and investment that has been proposed by Kelly [70], Breiman [20], Latané [82], and many others. In the language of investment, $q(X_t)$ is the factor by which capital grows during period $t$ and the objective is to maximize the asymptotic growth rate of the compounded capital $\prod_{0 \leq t < n} q(X_t)$. If $\{X_t\}$ is not an independent identically distributed sequence but a stationary process then we select $q(x_t|X^t)$ with knowledge of the past $X^t = (X_0, \ldots, X_{t-1})$ so as to maximize the growth rate of the compounded product $q(X^n) = \prod_{0 \leq t < n} q(x_t|X^t)$. With probability one, the maximum long run average growth rate is attained by always selecting the conditionally log-optimum $q^*(x_t|X^t)$. The asymptotic optimality of conditionally log-optimum portfolios has been discussed by many
authors including Breiman [20], Thorp [118], Finkelstein and Whitley [49], Móri and Székely [93], and Algoet and Cover [3]. The growth rate of compounded wealth has been called the “doubling rate” by Cover and Thomas [27], perhaps in analogy with the “half-life” of radioactive material. Log-optimum selections are also competitively optimum in the short run, by the game-theoretic arguments of Bell and Cover [11] [12]. We focus mostly on the independent identically case, but nearly every result in this paper can be generalized in one form or another to stationary processes.

Log-optimality is a natural concept in maximum likelihood estimation. In this context, $\mathcal{Q}$ is a model class of probability mass functions or probability densities and $P$ is the empirical distribution of some independent identically distributed samples. The computation of maximum growth exponents arises also in rate-distortion theory, albeit in a less explicit form. Let $X$ be a random variable with distribution $P$ on a source alphabet $\mathcal{X}$, let $\mathcal{Y}$ be a reproduction alphabet, and consider a distortion function $\rho(x, y) \geq 0$ on $\mathcal{X} \times \mathcal{Y}$. For any $s < 0$ consider the line with slope $s$ that supports the graph of the rate-distortion function $R(D)$. It turns out that the vertical intercept of this line is equal to $-W_{\mathcal{Q}_s}(P)$ where $\mathcal{Q}_s$ is the convex hull of the family $\{q^{y}_s(x)\}_{y \in \mathcal{Y}}$, where $q^{y}_s(x) = e^{\rho(x, y)}$. By varying the slope $s$ from 0 to $-\infty$ one obtains a representation of the convex decreasing function $R(D)$ as the hull of its supporting lines, namely $R(D) = \sup_{s \leq 0} [sD - W_{\mathcal{Q}_s}(P)]$.

In this paper we study the duality between the maximum growth exponent of functions in a convex family $\mathcal{Q}$ and the minimum information relative to measures in a convex family $\mathcal{M}$. First we recall the definition of information.

Let $m$ be a $\sigma$-finite reference measure on the space $\mathcal{X}$ and let $\mathcal{Q}(m)$ denote the family of densities $q(x)$ of $m$-dominated probability or subprobability measures $dQ = q dm$:

$$\mathcal{Q}(m) = \{q(x) : q(x) \geq 0, \int q dm \leq 1\}.$$

The growth exponent $E_P\{\log q(X)\}$ of the density $q = dQ/dm$ is equal to the exponential growth rate of the likelihood when independent random variables with distribution $P$ (selected by nature) are modeled according to a distribution $Q$ (selected by the statistician). Let $I_m(P)$ denote the maximum growth exponent of densities in the family $\mathcal{Q}(m)$:

$$I_m(P) = \sup_{q \in \mathcal{Q}(m)} E_P\{ \log q(X) \} = W_{\mathcal{Q}(m)}(P).$$

If $P$ is dominated by $m$ then the maximum growth exponent is attained by modeling according to the true distribution $P(dx)$ with density $p(x) = dP/dm(x)$. Thus the density $p(x)$ is log-optimum in the convex family $\mathcal{Q}(m)$ and $I_m(P)$ is equal to the Boltzmann-Gibbs-Shannon neg-entropy or information

$$I_m(P) = E_P\{ \log p(X) \}.$$

The family $\mathcal{Q}(m)$ is special because the log-optimum selection in $\mathcal{Q}(m)$ is explicitly known.

One may interpret $I_m(P)$ as the maximum growth rate of a gambler’s compounded wealth when gambling on independent realizations $X_t$ of the random variable $X$ against odds defined by the reference measure $m$. The optimum strategy is to bet according to the density $p$ of the true distribution $dP = p dm$ of the outcomes $X_t$. Betting according to the density $q$ of any alternative distribution $dQ = q dm$ will incur a loss in growth exponent equal to the relative entropy or Kullback-Leibler information divergence $I_Q(P) = I(P|Q)$. 

3
Let $\mathcal{M}$ be a family of $\sigma$-finite measures on $\mathcal{X}$. We say that a measure $m^* \in \mathcal{M}$ is log-optimum in $\mathcal{M}$ if $m^*$ attains the minimum information functional

$$I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} I_m(P).$$

If $\mathcal{M}$ is convex then the log-optimum $m^*(dx)$ in $\mathcal{M}$ is characterized by Karush-Kuhn-Tucker conditions. The measure $m^*(dx)$ need not be unique, but the likelihood $dP/dm^*(X)$ is uniquely defined up to almost sure equivalence. A bookmaker who can post odds according to any measure in the family $\mathcal{M}$ will minimize the exponential growth rate of his return payments to gamblers by setting the odds according to the log-optimum $m^*(dx)$ in $\mathcal{M}$.

For any family $\mathcal{M}$ of $\sigma$-finite measures $m(dx)$ on $\mathcal{X}$, we consider the polar family $\mathcal{Q}(\mathcal{M})$ of functions $q(x) \geq 0$ such that $\int q \ dm \leq 1$ for all $m \in \mathcal{M}$. If $\mathcal{M}$ is convex then the minimum information rate $I_\mathcal{M}(P)$ is equal to the maximum growth exponent $W_{\mathcal{Q}(\mathcal{M})}(P)$:

$$\inf_{m \in \mathcal{M}} I_m(P) = \sup_{q \in \mathcal{Q}(\mathcal{M})} W_q(P).$$

Dually, for any family $\mathcal{Q}$ of functions $q(x) \geq 0$ we consider the polar family $\mathcal{M}(\mathcal{Q})$ of measures $m(dx)$ such that $\int q \ dm \leq 1$ for all $q \in \mathcal{Q}$. If $\mathcal{Q}$ is convex then the maximum growth exponent $W_\mathcal{Q}(P)$ is equal to the minimum information functional $I_{\mathcal{M}(\mathcal{Q})}(P)$:

$$\sup_{q \in \mathcal{Q}} W_q(P) = \inf_{m \in \mathcal{M}(\mathcal{Q})} I_m(P).$$

If $\mathcal{Q}$ is convex and $\mathcal{M}$ is the polar of $\mathcal{Q}$ or vice versa, then $P(dx) = q^*(x) \ m^*(dx)$. Thus the log-optimum $q^*(x)$ in $\mathcal{Q}$ and the log-optimum $m^*(dx)$ in $\mathcal{M}$ are connected by the identity

$$q^*(X) = \frac{dP}{dm^*}(X) \text{ almost surely.}$$

The decomposition $P(dx) = q^*(x) \ m^*(dx)$ is similar to a factorization result of Lozanovskii [89] which is well known in the theory of Banach function spaces. In fact, the correspondence between $\mathcal{Q}$ and its polar $\mathcal{M}$ is essentially the correspondence between the unit ball for a function seminorm $\rho_\mathcal{Q}(q)$ and the unit ball for the conjugate seminorm $\rho_\mathcal{M}(m)$. The seminorms $\rho_\mathcal{Q}(q)$ and $\rho_\mathcal{M}(m)$ depend only on the absolute value of $q$ and $m$, so we may restrict our attention to nonnegative functions and measures when discussing polar sets.

Suppose a particular selection $\gamma(x)$ in the convex family $\mathcal{Q}$ is singled out as a reference for comparison. The maximum relativized growth exponent is defined as

$$W_{\mathcal{Q}/\gamma}(P) = \sup_{q \in \mathcal{Q}} E_P \left\{ \log \left( \frac{q(X)}{\gamma(X)} \right) \right\} = W_{\mathcal{Q}/\gamma}(X).$$

This is the loss in growth exponent when $\gamma$ rather than the log-optimum $q^*$ is selected in $\mathcal{Q}$. The polar family $\mathcal{M}(\mathcal{Q}/\gamma)$ is equal to the family $\gamma \mathcal{M}(\mathcal{Q})$ of subprobability measures $\gamma dm$ where $m \in \mathcal{M}(\mathcal{Q})$, and the minimax formula $W_{\mathcal{Q}/\gamma}(P) = I_{\mathcal{M}(\mathcal{Q}/\gamma)}(P)$ is always valid. In some cases, $I_{\mathcal{M}(\mathcal{Q}/\gamma)}(P)$ is equal to the minimum divergence of the true distribution $P$ relative to probability distributions in $\mathcal{M}(\mathcal{Q}/\gamma) = \gamma \mathcal{M}(\mathcal{Q})$. These are exactly the probability distributions under which $\gamma$ is log-optimum in $\mathcal{Q}$. More generally, the loss in achievable growth exponent when we are not allowed to make arbitrary selections in $\mathcal{Q}$ but are constrained to some convex subfamily $\Gamma$ will be denoted by $W_{\mathcal{Q}/\Gamma}(P)$. We shall prove
that $W_{Q/\Gamma}(P)$ is often equal to the minimum divergence of $P$ relative to distributions under which the log-optimum selection in $Q$ happens to fall in $\Gamma$. Thus the advantage in growth exponent if we may select in $Q$ rather than $\Gamma$ is equal to the minimum divergence of $P$ relative to distributions for which the extra flexibility provides no advantage at all.

The maximum growth exponent when the selections can be made in the convex family $Q$ with knowledge of side information $Y$ is defined as

$$W_Q(X|Y) = \sup_{q(x|y) \in Q} E_P\{\log q(X|Y)\}.$$  

The maximum is attained by some conditionally log-optimum $q^*(x|y)$ that is characterized by conditional Karush-Kuhn-Tucker conditions.

A transition kernel $\mu(dx|y)$ from a measurable space $Y$ to a measurable space $X$ is a measurable function from $Y$ to the space of measures on $X$. (The space of measures on $X$ is equipped with the smallest $\sigma$-field such that $B \mapsto m(B)$ is measurable in $m$ for any measurable set $B \subseteq X$.) We assume that $\mu(dx|y)B(dy)$ is a $\sigma$-finite measure on $X \times Y$ whenever $B(dy)$ is a probability measure on $Y$. If $X$ and $Y$ are random variables with joint distribution $P(dx\,dy) = P(dx|y)P(dy)$ on $X \times Y$, then the conditional information $I_\mu(X|Y)$ is defined as the information of $P(dx\,dy)$ relative to $\mu(dx|y)P(dy)$. For any convex family $\Pi$ of transition kernels $\mu(dx|y)$ from $Y$ to $X$ we define the minimum conditional information

$$I_\Pi(X|Y) = \inf_{\mu(dx|y) \in \Pi} I_\mu(X|Y).$$

A transition kernel $\mu^*(dx|y)$ in $\Pi$ which attains the minimum is conditionally log-optimum given $Y$ and is characterized by conditional Karush-Kuhn-Tucker conditions.

The advantage in achievable growth exponent for an agent who may select in $Q$ with knowledge of the side information $Y$ relative to an agent who must ignore $Y$ is denoted by

$$W_Q(X;Y) = W_Q(X|Y) - W_Q(X).$$

In the special case of gambling on $X$ against odds defined by $m(dx)$, the advantage in growth exponent is equal to the mutual information

$$I(X;Y) = I_m(X|Y) - I_m(X).$$

We shall prove that $W_Q(X;Y)$ is the minimum information divergence of $P(dx\,dy)$ relative to joint distributions under which the log-optimum selection in $Q$ given $Y$ is independent of $Y$. In particular, the loss in growth exponent when ignoring $Y$ is bounded by the mutual information $I(X;Y)$, which is the divergence of $P(dx\,dy)$ relative to $P(dx)P(dy)$. It turns out that the mutual information is decomposed into two complementary terms which express the value of the side information for two complementary purposes. Namely,

$$I(X;Y) = W_Q(X;Y) + W_M(X;Y),$$

where $M$ is the polar of $Q$ and $W_M(X;Y)$ is the advantage in information divergence that accrues if one may select a measure in the family $M$ with rather than without knowledge of the side information $Y$. Thus we get a deeper understanding of the inequality $W_Q(X;Y) \leq I(X;Y)$ which was first proved for log-optimum investment by Barron and Cover [10].
We are especially interested in convex families that are defined as the convex hull of some given family of extreme points. In particular, let $\gamma(x|y) \geq 0$ be a measurable function on $\mathcal{X} \times \mathcal{Y}$ and consider the convex family $\mathcal{Q}$ of mixtures

$$ q^B(x) = \int \gamma(x,y) B(dy), $$

where $B(dy)$ ranges over all probability measures on $\mathcal{Y}$. There is an iterative procedure which, given a suitable starting point $B_0(dy)$, generates a sequence of distributions $B_k(dy)$ such that the growth exponent of $q_k(x) = \int \gamma(x|y) B_k(dy)$ increases to the maximum:

$$ E_P\{\log q_k(X)\} \nearrow W_\mathcal{Q}(P) = E_P\{\log q^*(X)\}. $$

The sequence $\log q_k(X)$ will converge in $L^1(P)$ to $\log q^*(X)$ where $q^*$ is log-optimum in $\mathcal{Q}$.

Transforming $B_k(dy)$ into $B_{k+1}(dy)$ is essentially Cover's [25] iterative procedure for computing log-optimum portfolios. Each log-optimum $B^*(dy)$ will be a fixed point of the transformation. The transformation from $B_k(dy)$ to $B_{k+1}(dy)$ can be decomposed into two steps, first from $B_k(dy)$ to some transition kernel $F_k(dy|x)$ and then from $F_k(dy|x)$ to $B_{k+1}(dy)$. This decomposition shows that Cover's algorithm is related to Blahut's [17] alternating minimization algorithm for computing points on the rate-distortion curve of a memoryless source. To better see the connection we shall prove that

$$ W_\mathcal{Q}(P) = \sup_{F(dy|x)} [E_{PF}\{\log \gamma(X|Y)\} - I_{PF}(X;Y)] $$

where $I_{PF}(X;Y)$ denotes the mutual information when $X$ and $Y$ have joint distribution $P(dx)F(dy|x)$. If $s < 0$ and $\gamma(x|y) = e^{s\rho(x,y)}$ where $\rho(x,y) \geq 0$ is a distortion function, then the alternating minimization procedure converges to a point where the rate-distortion function $R(D)$ has slope $s$. Also, the EM algorithm for computing maximum likelihood estimates in mixture families is essentially the same as the Blahut algorithm. Alternating minimization algorithms are discussed in great generality and illustrated with many examples by Csiszár and Tusnády [38].

We now summarize the main point of each section in the paper. In Section II we discuss variational characterizations of the information or neg-entropy functional $I_m(P)$. By putting forward a variational definition for $I_m(P)$, we emphasize that the functional $I_m(P)$ is the maximum growth exponent $W_\mathcal{Q}(P)$ of some special convex family $\mathcal{Q}$. In Section III we discuss the existence and uniqueness of log-optimum selections in convex families and the Karush-Kuhn-Tucker conditions that uniquely characterize log-optimum selections up to almost sure equivalence. In Section IV we prove analogous results for log-optimum selections that achieve the minimum information relative to a convex family of measures. In Section V we introduce the concept of polar family and convex corner, and we prove that the maximum growth exponent $W_\mathcal{Q}(P)$ is equal to the minimum information $I_{\mathcal{M}}(P)$ if the families $\mathcal{Q}$ and $\mathcal{M}$ are polar of one another. In Section VI we study the decomposition of the mutual information $I(X;Y)$ into complementary terms that express the value of side information for complementary purposes. In Section VII we fix some measurable function $\gamma(x,y) \geq 0$ and we discuss an iterative procedure for computing the log-optimum $q^*(x) = \int \gamma(x|y) B^*(dy)$ among all mixtures $q^B(x) = \int \gamma(x,y) B(dy)$, where $B(dy)$ ranges over all probability distributions on $\mathcal{Y}$. Finally, in Section VIII we show how the concepts of log-optimality arise when maximizing determinants or volumes and when optimizing spectra. For basic concepts of convex analysis and convex duality we refer to the standard reference text of Rockafellar [108].
II. VARIATIONAL CHARACTERIZATIONS OF INFORMATION

The Boltzmann-Gibbs-Shannon neg-entropy of a probability distribution \( P(dx) \) with respect to a \( \sigma \)-finite reference measure \( m(dx) \) on a measurable space \( \mathcal{X} \) is usually defined as the expected log-likelihood \( I_m(P) = E_P[\log dP/dm(X)] \), provided the density \( p(x) = dP/dm(x) \) exists and \( \log p(X) \) is integrable. We define \( I_m(P) \) as the maximum growth exponent of densities \( q(x) = dQ/dm(x) \) of \( m \)-dominated probability or sub-probability measures \( Q(dx) = q(x) m(dx) \), and immediately verify that the maximum is attained by the true density \( p(x) = dP/dm(x) \). Thus we view \( I_m(P) \) as the maximum growth exponent of a special convex family. The variational formula reveals a natural gambling interpretation for \( I_m(P) \), and it enables us to interpret \( I_m(P) \) as a maximum expected utility function. The well known variational characterizations of Kullback-Leibler divergence in Pinsker’s monograph [101] and in textbooks on large deviation theory are derived with little effort.

A. Information as maximum growth exponent

Let \( m(dx) \) be a \( \sigma \)-finite reference measure on a measurable space \( \mathcal{X} \) and let \( X \) be a random variable with distribution \( P \) on \( \mathcal{X} \). Let \( Q(m) \) be the convex family of measurable functions \( q(x) \geq 0 \) such that \( \int q \, dm \leq 1 \). The maximum growth exponent of selections in \( Q(m) \) will be called the information of \( P \) relative to \( m \) and will be denoted by

\[
I_m(P) = \sup_{Q \in Q(m)} E_P[\log q(X)] = W_{Q(m)}(P). \tag{1}
\]

It suffices to take the supremum over those \( q(x) \) in \( Q(m) \) such that \( \int q \, dm = 1 \). Indeed, if \( 0 < \int q \, dm < 1 \) then the normalized function \( q(x)/\int q \, dm \) also belongs to \( Q(m) \), and its growth exponent exceeds that of \( q(x) \) by the positive amount \( -\log(\int q \, dm) \).

Observe that \( dQ = q \, dm \) is a subprobability measure if \( \int q \, dm \leq 1 \), and a probability measure if \( \int q \, dm = 1 \). Thus (1) asserts that

\[
I_m(P) = \sup_Q E_P \left\{ \log \left( \frac{dQ}{dm}(X) \right) \right\}, \tag{2}
\]

where the supremum is taken over \( m \)-dominated sub-probability or probability measures \( dQ = q \, dm \) with densities \( q = dQ/dm \). It is easy to see that the maximum is attained by the true distribution \( P \) when \( P \) admits a density \( p = dP/dm \).

**Theorem 1.** If \( P \) is dominated by \( m \), then the density \( p = dP/dm \) is log-optimum in the convex family \( Q(m) \). If the expectation \( E_P[\log p(X)] \) is well defined, then

\[
I_m(P) = E_P[\log p(X)] = E_P \left\{ \log \left( \frac{dP}{dm}(X) \right) \right\}. \tag{3}
\]

If \( P \) is not dominated by \( m \) then \( I_m(P) = \pm \infty \), depending on whether or not there exists some \( q \in Q(m) \) such that \( E_P[\log q(X)] > -\infty \).

**Proof:** If \( P \) is dominated by \( m \) then the density \( p = dP/dm \) is a function in \( Q(m) \), since \( \int p \, dm = P(\mathcal{X}) = 1 \). For any \( q \in Q(m) \), we have

\[
E_P \left\{ \frac{q(X)}{p(X)} \right\} = \int \frac{q}{p} \, p \, dm = \int_{\{x : p(x) > 0\}} q \, dm \leq \int q \, dm \leq 1,
\]

7
and consequently, by Jensen’s inequality,

\[ E_P \left\{ \log \left( \frac{q(X)}{p(X)} \right) \right\} \leq \log \left( E_P \left\{ \frac{q(X)}{p(X)} \right\} \right) \leq \log 1 = 0. \]

It follows that \( p \) is log-optimum in \( Q(m) \). If \( E_P \{ \log q(X) \} \) is well defined then \( p(x) \) is a function in \( Q(m) \) attaining the maximum growth exponent \( I_m(P) = W_{Q(m)}(P) \).

If \( P \) is not dominated by \( m \), then there exists a measurable subset \( B \) of \( X \) such that \( P(B) > 0 \) and \( m(B) = 0 \). Suppose \( E_P \{ \log q(X) \} > -\infty \) for some \( q \in Q(m) \), and for \( a \geq 0 \) let \( q_a(x) = q(x) + a1_B(x) \) where \( 1_B(x) \) is the indicator function of \( B \). Then \( q_a(x) \) is a function in \( Q(m) \) and \( E_P \{ \log q_a(X) \} \) increases to \( I_m(P) = \infty \) as \( a \to \infty \) by the monotone convergence theorem. This suffices to prove the theorem.

If \( m \) is Lebesgue measure on the real line then \( I_m(P) = E_P \{ \log p(X) \} \) is minus the differential entropy of \( P \). If \( m \) is counting measure on a finite or countable set, then \( I_m(P) \) is minus the Shannon entropy. Shannon [113] argued that a discrete random variable \( X \) with probability mass function \( P(x) \) is best described on average by a code with length \( -\log P(X) \), rounded up to the nearest integer. Thus \( -\log P(X) \) is the ideal code-word length or intrinsic complexity of \( X \), and \( \log[P(X)/Q(X)] \) may be interpreted as the pointwise redundancy of an alternative model \( Q \). The expected redundancy or increase in average per-symbol description length for a code that is tailored to \( Q \) rather than the true distribution \( P \) is equal to the information divergence \( I(P|Q) = E_P \{ \log[dP/dQ(X)] \} \).

Suppose \( m \) is a finite measure with total mass \( m(X) < \infty \). The constant \( \gamma(x) = 1/m(X) \) belongs to \( Q(m) \), and the inequality \( I_m(P) \geq E_P \{ \log \gamma(X) \} \) reduces to

\[ I_m(P) \geq -\log m(X) > -\infty. \]

In particular, \( I_m(P) \geq 0 \) if \( m \) is subnormalized. If \( P \) and \( Q \) are probability measures with densities \( p = dP/dm \) and \( q = dQ/dm \), then the difference between \( E_P \{ \log q(X) \} \) and the maximum growth exponent \( E_P \{ \log p(X) \} \) is equal to the information divergence

\[ I_Q(P) = I(P|Q) = E_P \left\{ \log \left( \frac{dP}{dQ}(X) \right) \right\} = E_P \left\{ \log \left( \frac{p(X)}{q(X)} \right) \right\}. \]

The number \( I(P|Q) \) was introduced by Kullback and Leibler [81] and is called the relative entropy, cross-entropy, directed divergence, or information for discriminating \( P \) from \( Q \). Sometimes \( I(P|Q) \) is called the Kullback-Leibler distance because \( I(P|Q) \geq 0 \) with equality iff \( P = Q \). However, \( I(P|Q) \) lacks symmetry and does not satisfy the triangle inequality, so it is not a metric. The inequality \( I(P|Q) \geq 0 \) or \( E_P \{ \log q(X) \} \leq E_P \{ \log p(X) \} \), which is sometimes attributed to Gibbs, follows immediately from Klein’s inequality \( \log(q/p) \leq (q/p) - 1 \), by integrating both sides with respect to \( dp = p \, dm \). Wald [126] derived it from Jensen’s inequality and used it to prove the consistency of maximum likelihood parameter estimates. The statistical interpretation of entropy and relative entropy can be traced back to work of Boltzmann in statistical mechanics, cf. Akaike [1].

The concept of information also arose in the work of Good [58] on rational decisions and work of Kelly [70] on gambling. Essentially, Theorem 1 states that the function \( u_m(Q, x) = \log[dQ/dm(x)] \) is a proper scoring rule in the sense of de Finetti and Savage [111]. Proper scoring rules are incentives that induce a consultant to reveal her subjective beliefs about the distribution of a random variable \( X \). Suppose the consultant is promised a reward
\( u(Q, x) \) if she proposes \( Q \) as forecast distribution and the random variable \( X \) then takes on the actual value \( x \). A rational consultant whose beliefs are expressed by a subjective distribution \( P \) will choose \( Q \) so as to maximize the expected reward \( E_P\{u(Q, X)\} \). A scoring rule \( u(Q, x) \) is called proper if the maximum expected reward is attained by the subjective distribution \( P \), i.e. if \( E_P\{u(Q, X)\} \leq E_P\{u(P, X)\} \) for all \( Q \) and hence

\[
\sup_Q E_P\{u(Q, X)\} = E_P\{u(P, X)\}.
\]

The loss in expected score if the consultant proposes a distribution \( Q \) other than her subjective distribution \( P \) is equal to the divergence

\[
D(P|Q) = E_P\{u(P, X)\} - E_P\{u(Q, X)\}.
\]

A proper scoring rule is called strictly proper if the inequality \( D(P|Q) \geq 0 \) holds with strict equality only when \( Q = P \). The logarithmic scoring rule \( u_m(Q, x) = \log[dQ/dm(x)] \) is strictly proper, and its divergence \( D(P|Q) \) is equal to the Kullback-Leibler divergence \( I(P|Q) \). Bernardo [15] and others have shown that if \( ||\mathcal{X}|| > 2 \) then except for the choice of \( m(dx) \) and a constant multiplicative factor, the logarithmic scoring rule \( u_m(Q, x) = \log[dQ/dm(x)] \) is the unique proper scoring rule for which the score \( u(Q, x) \) depends on \( Q \) and \( x \) only though the density \( q(x) = dQ/dm(x) \) at the actual observation \( x \). The choice of the reference measure \( m(dx) \) is fairly arbitrary because what really matters are comparisons which involve only score differences \( u_m(Q, x) - u_m(Q', x) = \log[dQ/dQ'(x)] \).

### B. Information and quantization

The information in a random variable \( X \) is equal to the maximum information contained in quantized or coarse-grained versions of \( X \). There are many proofs of this result in the literature. We give an elementary proof based on the definition of the Lebesgue integral as the maximum integral of simple approximations from below.

A measurable function on \( \mathcal{X} \) is called simple if it takes on only finitely many values. If the sets where the function takes on its distinct values belong to a field \( \mathcal{A} \), then the function is called \( \mathcal{A} \)-measurable. We need the following lemma, which establishes lower semicontinuity of the integral \( \int V \, dP \) in \( P \) if the integrand \( V(x) \) is bounded below.

**Lemma 1.** Let \( X \) be a random variable with distribution \( P(dx) \) on a measurable space \((\mathcal{X}, \mathcal{B})\) and let \( \mathcal{A} \) be a generating subfield of the \( \sigma \)-field \( \mathcal{B} \). If a measurable function \( V(x) \) is bounded below on \( \mathcal{X} \), then \( \int V \, dP = \sup \int U \, dP \), where the supremum is taken over simple \( \mathcal{A} \)-measurable functions \( U(x) \) such that \( U(x) \leq V(x) \).

**Proof:** We may assume without loss of generality that \( V(x) \geq 0 \). If \( \int V \, dP < \infty \), then for any \( \epsilon > 0 \) there exists a simple \( \mathcal{B} \)-measurable function \( V'(x) = \sum_{1 \leq k \leq K} a_k \mathbb{1}\{x \in B_k\} \) such that \( V'(x) \leq V(x) \) and \( \int V' \, dP \geq \int V \, dP - \epsilon \). Without loss of generality, \( \{B_k\}_{1 \leq k \leq K} \) is a disjoint family and \( a_k \geq 0 \) for all \( k \). Let \( a = \sum_{1 \leq k \leq K} a_k \), and choose sets \( A_k \in \mathcal{A} \) such that \( A_k \subseteq B_k \) and \( P(B_k \setminus A_k) \leq \epsilon/(K a) \). Then \( U(x) = \sum_{1 \leq k \leq K} a_k \mathbb{1}\{x \in A_k\} \) is a simple \( \mathcal{A} \)-measurable function such that \( U(x) \leq V'(x) \leq V(x) \) and \( \int U \, dP \geq \int V' \, dP - \epsilon \geq \int V \, dP - 2\epsilon \). If \( \int V \, dP = \infty \), we can similarly construct a simple \( \mathcal{A} \)-measurable function \( U(x) \) such that \( U(x) \leq V(x) \) and \( \int U \, dP \geq \epsilon^{-1} \).

Given a measurable space \((\mathcal{X}, \mathcal{B})\), let \( \Pi^c(\mathcal{B}) \) denote the class of all partitions of \( \mathcal{X} \) into a countable collection of measurable subsets. For any partition \( \beta \in \Pi^c(\mathcal{B}) \) and \( x \in \mathcal{X} \) let
\( \beta(x) \) denote the atom of \( \beta \) that happens to contain the point \( x \) and let

\[
p_\beta(x) = \frac{P(\beta(x))}{m(\beta(x))},
\]

\[
I_m^\beta(P) = E_P\{\log p_\beta(X)\} = \sum_{B \in \text{Atoms}(\beta)} P(B) \log \left( \frac{P(B)}{m(B)} \right).
\]

If \( \mathcal{A} \) is an algebra or field of measurable subsets of \( \mathcal{X} \), then \( \Pi(\mathcal{A}) \) will denote the set of all partitions \( \alpha \) of \( \mathcal{X} \) into a finite collection of sets in \( \mathcal{A} \) (such sets are called \( \mathcal{A} \)-measurable).

**Theorem 2.** Let \( P \) be a probability measure and let \( m \) be a \( \sigma \)-finite reference measure on a measurable space \( (\mathcal{X}, \mathcal{B}) \). The supremum \( I_m(P) = \sup_{q \in \mathcal{Q}(m)} E_P\{\log q(X)\} \) may be taken over the elementary functions in \( \mathcal{Q}(m) \):

\[
I_m(P) = \sup_{\beta \in \Pi(\mathcal{B})} I_m^\beta(P).
\]

If \( m \) is a finite measure and \( \mathcal{A} \) is a generating subfield of \( \mathcal{B} \), then it suffices to take the supremum over simple \( \mathcal{A} \)-measurable functions in \( \mathcal{Q}(m) \):

\[
I_m(P) = \sup_{\alpha \in \Pi(\mathcal{A})} I_m^\alpha(P). \tag{4}
\]

**Proof:** If \( \beta \) is a measurable partition of \( \mathcal{X} \) then \( p_\beta(x) \) is a function in \( \mathcal{Q}(m) \) and hence

\[
I_m^\beta(P) = E_P\{\log p_\beta(X)\} \leq I_m(P) = \sup_{q \in \mathcal{Q}(m)} E_P\{\log q(X)\}.
\]

Consider an arbitrary function \( q \in \mathcal{Q}(m) \) such that \( E_P\{\log q(X)\} > -\infty \). For \( k \geq 1 \) let \( \beta_k \) denote the partition of \( \mathcal{X} \) with atoms

\[
B_k^i = \{ x : i2^{-k} \leq \log q(x) < (i + 1)2^{-k} \}, \quad -\infty < i < \infty
\]

and let

\[
q_k(x) = \sum_i \exp(i2^{-k})1\{ x \in B_k^i \}.
\]

Note that \( q_k \in \mathcal{Q}(m) \) since \( q_k \leq q \) and \( q \in \mathcal{Q}(m) \). Since \( E_P\{\log q_k(X)\} \leq I_m^\beta_k(P) \) and \( E_P\{\log q_k(X)\} \) increases to \( E_P\{\log q(X)\} \) by the monotone convergence theorem, we have

\[
E_P\{\log q(X)\} \leq \sup_{\beta \in \Pi(\mathcal{B})} I_m^\beta(P).
\]

Since this is true for arbitrary \( q \in \mathcal{Q}(m) \) such that \( E_P\{\log q(X)\} > -\infty \), we may conclude that the following chain of inequalities holds with equality throughout:

\[
I_m(P) = \sup_{q \in \mathcal{Q}(m)} E_P\{\log q(X)\} \leq \sup_{\beta \in \Pi(\mathcal{B})} I_m^\beta(P) \leq I_m(P).
\]

If \( m \) is a finite measure then the constant \( \gamma(x) = 1/m(\mathcal{X}) \) is a function in \( \mathcal{Q}(m) \). If \( q \in \mathcal{Q}(m) \) and \( 0 < \epsilon < 1 \) then the convex combination \( q_\epsilon(x) = (1 - \epsilon)\gamma(x) + \epsilon q(x) \) is
a function in $Q(m)$. The function $\log q_\epsilon(x)$ is bounded below since $q_\epsilon(x) \geq (1 - \epsilon)\gamma(x)$, and its growth exponent $E_P\{\log q_\epsilon(X)\}$ is bounded below by $E_P\{\log q(X)\} + \log \epsilon$ since $q_\epsilon(x) \geq \epsilon q(x)$. Letting $\epsilon \to 1$, we see that $\sup_{0 < \epsilon < 1} E_P\{\log q_\epsilon(X)\} \geq E_P\{\log q(X)\}$ and the supremum $I_m(P) = \sup_{q \in Q(m)} E_P\{\log q(X)\}$ may be taken over functions in $Q(m)$ whose logarithm is bounded below. If $\log q(x)$ is bounded below then by Lemma 1, $E_P\{\log q(X)\}$ is equal to the maximum growth exponent of simple $\mathcal{A}$-measurable functions below $q(x)$. Thus $I_m(P)$ is a supremum of suprema and is equal to the maximum growth exponent of functions $q(x)$ in $Q(m)$ such that $\log q(x)$ is constant on the atoms of some finite $\mathcal{A}$-measurable partition of $\mathcal{X}$. But the growth exponent $E_P\{\log q(X)\}$ of any function $q(x)$ in $Q(m)$ that is constant on the atoms of some finite partition $\alpha$ is not larger than $I_{\alpha}^\alpha(P) = E_P\{\log p_{\alpha}(X)\}$ since $p_\alpha(x)$ is log-optimum in the convex set of functions in $Q(m)$ that are constant on each atom of $\alpha$. We may conclude that $I_m(P) = \sup_{\alpha} E_P\{\log p_{\alpha}(X)\}$ where the supremum is taken over all finite partitions $\alpha$ with atoms in $\mathcal{A}$. This proves (4).

The classes $\Pi^c(\mathcal{B})$ and $\Pi(\mathcal{A})$ are directed with respect to refinement of partitions. The supremum of $I_{\alpha}^\alpha(P)$ over either class may be taken over a cofinal subclass because $I_{\alpha}^\alpha(P)$ can only increase when partition $\beta$ is refined. If $P$ and $M$ are probability measures on a separable metric space $\mathcal{X}$ and $\{\alpha_k\}_{k \geq 1}$ is an increasing sequence of finite partitions that asymptotically generate the Borel $\sigma$-field on $\mathcal{X}$, then

$$I_{\alpha_k}^\alpha(P) \nearrow I_m(P) \quad \text{as } k \to \infty.$$  

The functional $I_m(P)$ represents the information in the random variable $X$ and may be denoted informally by $I_m(X)$. Similarly, $I_{\alpha_k}^\alpha(P)$ represents the information in the quantized random variable $\beta(X)$ and may be denoted informally by $I_m(\beta(X))$. The mutual information $I(X; Y)$ between two random variables $X$ and $Y$ is a special case of information divergence, namely the divergence between the joint distribution $P(dx \, dy)$ and the product of the marginals $P(dx) \, P(dy)$. Gelfand, Kolmogorov and Yaglom [53-54] proved that

$$I(X; Y) = \sup_{\alpha, \beta} I(\alpha(X); \beta(Y)),$$

where the supremum is taken over finite measurable partitions $\alpha$ of $\mathcal{X}$ and $\beta$ of $\mathcal{Y}$. This follows from (4) because the $\sigma$-field on $\mathcal{X} \times \mathcal{Y}$ is generated by measurable rectangles.

Our proof that $I_m(P) = \sup_{\alpha \in \Pi(\mathcal{A})} I_{\alpha}^\alpha(P)$ is a simplification of many proofs in the literature. We observed that $I_m(P)$ is the maximum Lebesgue integral of functions in a certain family, and each Lebesgue integral is the maximum integral of simple approximations from below. Ghurye [55] and Čencov [29] approximate the integral from both sides, so they depend on the Darboux-Young approach to integration that is described in Exercise 29 on p. 144 of Loevé [88]. Many authors including Perea [98], Hájek [63], Kallianpur [69], Moy [94], and Barron [8] rely on martingale properties. The martingale argument of Hájek [63] is very elegant, and so is the approach suggested in Exercises IV-5-3 and IV-5-4 of Neveu [95]. The proof of Dobrushin [44] [45] relies on a nontrivial combinatorial argument which is also presented in the texts of Pinsker [101] and Gray [59]. Guiașu [62] gives an elementary but somewhat technical proof based only on the monotone and dominated convergence theorems, while Kullback and Thall [78] and Kullback, Keegel and Kullback [79] present information-theoretic proofs that are very interesting but somewhat involved. All these authors assume that the reference measure is normalized, but Ghurye [55] and Csiszár [31] have considered the case where $m$ is $\sigma$-finite. Csiszár [32] has also formulated a
sufficient condition for nonmonotonic sequences of partitions \( \{ \beta_k \} \) that ensures convergence of \( I_m^\beta_k(P) \) to \( I_m(P) \) and has pointed out connections between \( I_m(P) \) and Rényi's concept of dimensional entropy. Masani [91] has removed the assumption that \( m \) is \( \sigma \)-finite.

Again let \( \mathcal{A} \) be a generating subfield of the \( \sigma \)-field on the space \( \mathcal{X} \). If \( m \) is a finite measure on \( \mathcal{X} \) then by Theorem 2, the supremum \( I_m(P) = \sup_{q \in \mathcal{Q}(m)} E_P\{ \log q(X) \} \) may be taken over simple \( \mathcal{A} \)-measurable densities in \( \mathcal{Q}(m) \). It suffices to consider densities with rational values since the rationals are dense in the reals. If \( \mathcal{A} \) is countable then by enumerating the countable set of all simple \( \mathcal{A} \)-measurable densities with rational values into a sequential list, we obtain the first part of the following theorem.

**Theorem 3.** Let \( m \) be a reference measure on a measurable space \( \mathcal{X} \) with a countably generated \( \sigma \)-field. If \( m \) is a finite measure then there exists a sequence of densities \( q_k(x) \) in \( \mathcal{Q}(m) \) such that for every probability measure \( P \) on \( \mathcal{X} \),

\[
I_m(P) = \sup_k E_P\{ \log q_k(X) \}.
\]  

(5)

If \( m(dx) \) is \( \sigma \)-finite and \( \gamma(x) \) is a particular density in \( \mathcal{Q}(m) \), then there exists a sequence \( \{ q_k(x) \} \) in \( \mathcal{Q}(m) \) such that (5) holds for all distributions \( P \) such that \( E_P\{ \log \gamma(X) \} > -\infty \).

**Proof:** The first part is already clear. To prove the second part, consider the sub-probability measure \( M(dx) = \gamma(x) m(dx) \) and observe that if \( E_P\{ \log \gamma(X) \} > -\infty \), then

\[
I_m(P) = E_P\{ \log \gamma(X) \} + I_M(P).
\]

If \( \mathcal{A} \) is a countable generating subfield then by the first part of the theorem, there exists a sequence of simple \( \mathcal{A} \)-measurable densities \( g_k(x) \) in \( \mathcal{Q}(M) \) such that

\[
I_M(P) = \sup_k E_P\{ \log g_k(X) \}.
\]

The second part of the theorem now follows by setting \( q_k(x) = \gamma(x) g_k(x) \), for all \( k \).

Theorem 3 proves the existence of a sequence of probabilistic models \( Q_k(dx) = q_k(x) m(dx) \) such that for every probability distribution \( P \) with \( E_P\{ \log \gamma(X) \} > -\infty \),

\[
I_m(P) = \sup_k E_P\left\{ \log \left( \frac{dQ_k}{dm}(X) \right) \right\}.
\]

**C. Gambling interpretation of \( I_m(P) \)**

We consider gambling on independent realizations \( X_t \) of a random variable \( X \) with values in the space \( \mathcal{X} \) against odds that are defined by a \( \sigma \)-finite reference measure \( m(dx) \) on \( \mathcal{X} \). To play one round of the betting game, the gambler must specify a partition \( \beta \) of \( \mathcal{X} \) and invest a nonnegative amount in every atom of this partition. When the random outcome \( X \) is revealed, the bookie will return the amount allocated to the atom \( \beta(X) \) that happens to contain \( X \), multiplied by the odds \( 1/m(\beta(X)) \) of that atom. Thus the return per invested unit is given by \( Q(\beta(X))/m(\beta(X)) \) if the gambler bets according to strategy \( Q \) on \( \mathcal{X} \), that is if he allocates a portion \( Q(B) \) of his total wealth to each atom \( B \) of \( \beta \).

Suppose a gambler starts with one unit of initial wealth and he redistributes his compounded wealth at the beginning of every round \( t \) according to some fixed subprobability
measure $Q(dx)$. If the random outcomes $X_i$ is always quantized according to partition $\beta$ then the compounded wealth after $n$ rounds amounts to

$$S_n = \prod_{0 \leq i < n} \frac{Q(\beta(X_i))}{m(\beta(X_i))}.$$ 

The strong law of large numbers for products asserts that $S_n$ grows exponentially fast almost surely with constant limiting rate

$$E_P \left\{ \log \left( \frac{Q(\beta(X))}{m(\beta(X))} \right) \right\} = \sum_{B \in \text{Atoms}(\beta)} P(B) \log \left( \frac{Q(B)}{m(B)} \right) = I_m(\beta(X)) - I_Q(\beta(X)).$$

Recall that $I_Q(\beta(X)) \geq 0$ with equality iff $Q$ and $P$ have the same restriction to the $\sigma$-field generated by $\beta$. Thus the maximum growth exponent of wealth when $X$ is quantized according to $\beta$ is equal to $I_m(\beta(X))$, and the maximum is attained by ignoring the odds measure $m$ and betting according to the true distribution $P$ of $X$. It is in the interest of the gambler to insist on as fine a partition $\beta$ of $X$ as is practically feasible, since $I_m(\beta(X))$ can only increase when $\beta$ is refined. In fact, if partition $\beta$ is refined and asymptotically generates the $\sigma$-field on $X$ then the growth exponent $I_m(\beta(X))$ increases to the information $I_m(\beta(X))$, and the loss in growth exponent when gambling according to an alternative measure $Q$ increases to the relative entropy $I_Q(X)$. We summarize our conclusions as follows.

**Theorem 4.** Suppose a gambler is allowed to bet on independent realizations $X_i$ of the random variable $X$, against odds that are defined by the reference measure $m(dx)$. The maximum growth rate of the gambler's compounded wealth is equal to the information functional $I_m(P)$ and is attained by betting according to the true distribution $P$ of $X$. If the gambler bets according to an alternative probability or sub-probability measure $Q$ on $X$, then he will incur a loss in growth exponent equal to the relative entropy $I_Q(P)$.

Suppose $P$ is dominated by $m$. If the gambler bets according to a sub-probability measure $Q$ with density $q = dQ/dm$ relative to $m$ and if $\beta$ is refined along a sequence of partitions $\beta_k$ that asymptotically generates the $\sigma$-field on $X$, then by the martingale convergence theorem, the return per invested unit converges to

$$q(X) = \frac{dQ}{dm}(X) = \lim_k \frac{Q(\beta_k(X))}{m(\beta_k(X))}$$

almost surely.

Gambling according to the density $p = dP/dm$ of the true distribution $P$ attains the maximum growth exponent $I_m(P)$, and betting according to the density $q = dQ/dm$ of the alternative measure $Q$ incurs a loss in growth exponent equal to the relative entropy $I_Q(P)$. Indeed, the growth exponent of wealth under strategy $dQ = q dm$ is given by

$$E_P \{ \log q(X) \} = E_P \{ \log p(X) \} - E_P \left\{ \log \left( \frac{p(X)}{q(X)} \right) \right\} = I_m(P) - I_Q(P)$$

and $I_Q(P) \geq 0$ with equality iff $Q = P$.

Suppose the gambler re-invests only a portion $\lambda < 1$ of his wealth according to some normalized distribution $M$. Such a betting strategy is concisely represented by the sub-probability measure $dQ = \lambda dM$ with total mass $Q(X) = \lambda$. The strategy $Q$ cannot be
log-optimum since a portion $1 - \lambda$ of wealth is wasted in every round. In fact, if all wealth is distributed according to the normalized measure $M(dx) = Q(dx)/Q(\mathcal{X})$ instead, then the growth exponent of wealth will improve by a positive amount $-\log Q(\mathcal{X})$.

Suppose two bookmakers post different odds measures $m_0(dx)$ and $m_1(dx)$. A gambler who wants to maximize the growth exponent of compounded wealth will distribute his wealth according to the true distribution $P(dx)$ on the space of possible outcomes $\mathcal{X}$ and will then assign the portion allocated to each outcome $x$ to the bookmaker who is posting the highest odds for that particular outcome $x$. The maximum growth rate of the compounded capital is the same as when betting against odds defined by the measure $m_0(dx) \land m_1(dx)$. Thus the maximum growth rate is equal to $E\{\log q^*(X)\}$ where

$$q^*(x) = \frac{dP}{dm_0 \land dm_1}(x) = \max \left[ \frac{dP}{dm_0}(x), \frac{dP}{dm_1}(x) \right].$$

D. Information as a convex conjugate

Let $X$ be a random variable with distribution $P$ and let $m$ be a $\sigma$-finite reference measure on a measurable space $\mathcal{X}$. The variational formulas for $I_m(P)$ in Section A can be written in the equivalent form

$$I_m(P) = \sup_{q(x) \geq 0, \int q dm = 1} E_P\{\log q(X)\}$$

$$= \sup_{u(x) \geq 0} E_P \left\{ \log \left( \frac{u(X)}{\int u dm} \right) \right\}$$

$$= \sup_{V(x)} E_P \left\{ \log \left( \frac{e^{V(X)}}{\int e^V dm} \right) \right\},$$

where the supremum is taken over functions $u(x) = e^{V(x)}$ such that $0 < \int u dm < \infty$. If $P$ is dominated by $m$ then by Theorem 1, the supremum is attained when $q(x)$ or $u(x)$ is equal to the density $p(x) = dP/dm(x)$ or when $V(x) = \log p(x)$. We may also write

$$I_m(P) = \sup_{u(x) > 0} \left[ \int \log u dP - \log(\int u dm) \right]$$

$$= \sup_{V(x)} \left[ \int \log u dP - \log(\int e^V dm) \right],$$

where the supremum is taken over functions $V(x) = \log u(x)$ that are upper bounded by a finite constant, so that the expectation $\int V dP = E_P\{\log u(X)\}$ is well defined and finite.

For any measurable function $V(x)$ on $\mathcal{X}$ let

$$\Lambda_m(V) = \log(\int e^V dm).$$

The functional $\Lambda_m(V)$ is convex in $V$ by Hölder's inequality. If $m$ is a finite measure then $I_m(P)$ and $\Lambda_m(V)$ are lower semicontinuous and convex conjugate functionals in the sense of Fenchel and Young. The fact that $I_m(P)$ is the convex conjugate of $\Lambda_m(V)$ is essentially Theorem 1, and the fact that $\Lambda_m(V)$ is the convex conjugate of $I_m(P)$ is essentially the minimum discrimination information theorem of Kullback and Khairat [80]. See also Section 17.2 of Guiaşu [62] and pp. 18–28 of Kullback et al. [79]. If $m$ is merely $\sigma$-finite, then
$I_m(P)$ and $\Lambda_m(V)$ are still convex conjugates, although perhaps not lower semicontinuous. The "inner product" of a function $V(x)$ and a measure $P(dx)$ is defined as the integral

$$(P,V) = \int V \, dP.$$ 

**Theorem 5.** Let $m$ be a $\sigma$-finite measure on a measurable space $X$. The functionals $I_m(P)$ and $\Lambda_m(V)$ are convex conjugates in the following sense.

(a) For any probability distribution $P(dx)$ on $X$ such that $I_m(P) > -\infty$, we have

$$I_m(P) = \sup_V [(P,V) - \Lambda_m(V)],$$

where the supremum is taken over functions $V(x)$ such that $\Lambda_m(V)$ is finite. If $P$ is dominated by $m$ then the supremum is attained by the log-likelihood $V_P(x) = \log[dP/dm(x)]$.

(b) For any measurable function $V(x)$ on $X$ we have

$$\Lambda_m(V) = \sup_P [(P,V) - I_m(P)],$$  

(7)

where the supremum is taken over all distributions $P(dx)$ and $[(P,V) - I_m(P)]$ is interpreted as minus the information of $dP$ relative to the measure $e^V \, dm$. If $\Lambda_m(V)$ is finite then the supremum is attained by the distribution

$$dP_V = \frac{e^V \, dm}{\int e^V \, dm} = e^{V-\Lambda_m(V)} \, dm.$$

**Proof:** Part (a) is a reformulation of Theorem 1 or the variational formula (6). To prove (b), observe that for any distribution $P(dx)$ such that $[(P,V) - I_m(P)]$ is well defined,

$$\Lambda_m(V) \geq [(P,V) - I_m(P)] = [-I_{e^V \, dm}(P)].$$

If $\Lambda_m(V)$ is finite then $P_V$ is a well defined probability distribution and equality holds in (7) for $P = P_V$. If $\Lambda_m(V) = -\infty$ then $e^V \, dm$ is the zero measure and $I_{e^V \, dm}(P) = \infty$. If $\Lambda_m(V) = \infty$ then for $k \geq 1$ we consider the distribution

$$dP_k = \frac{e^{(V+W) \wedge k} \, dM}{\int e^{(V+W) \wedge k} \, dM},$$

where $W(x)$ is a function and $M(dx)$ is a normalized distribution on $X$ such that $dm = e^W \, dM$. Clearly

$$-I_{e^V \, dm}(P_k) = \Lambda_M([(V + W) \wedge k] + E_{P_k} \left\{ \log \left( \frac{e^{V+W}}{e^{(V+W) \wedge k}} \right) \right\} \geq \Lambda_M([(V + W) \wedge k)$$

and $\Lambda_M([(V + W) \wedge k)$ tends to $\Lambda_M(V + W) = \Lambda_m(V) = \infty$ as $k \to \infty$. It follows that $\Lambda_m(V) = \sup_P [-I_{e^V \, dm}(P)]$, which is the desired conclusion (7).

Theorem 5 asserts that the graph of the convex functional $I_m(P)$ is the hull of its supporting hyperplanes. These supporting hyperplanes can be parametrized either in terms of their slope $V$ or in terms of their contact point $Q$. If $\Lambda_m(V)$ is finite then the functional $[(P,V) - \Lambda_m(V)]$ is affine in $P$ and its graph is the hyperplane with slope $V$ that supports the
graph of $I_m(P)$ at the unique contact point $P_V$. If $q(x) \geq 0$ is a function such that $\int q \, dm = 1$, then the graph of the linear functional $E_P\{\log q(X)\}$ is a hyperplane that supports the graph of $I_m(P)$ at the unique contact point $dQ = q \, dm$. The slope of the supporting hyperplane at $Q$ is equal to $V_Q = \log q$ where $q = dQ/dm$. On the other hand, if $\int q \, dm < 1$ then the graph of the linear functional $E_P\{\log q(X)\}$ is a hyperplane entirely below the graph of $I_m(P)$. This hyperplane must be shifted upward by the amount $-\Lambda_m(\log q) > 0$ until it supports the graph of $I_m(P)$. Contact will be made at the distribution $dM = q \, dm / (\int q \, dm)$ that is obtained by normalizing the subprobability $dQ = q \, dm$. Note that $\Lambda_m(\log q) = \log(\int q \, dm)$ is the logarithm of the normalizing factor.

Suppose $P$ and $M$ are probability measures on a metric space $\mathcal{X}$ with its Borel $\sigma$-field. Donsker and Varadhan [46] proved that the supremum $I_M(P) = \sup_{V(x)} [(P, V) - \Lambda_M(V)]$ may be taken over bounded continuous functions $V(x)$. The proof also appears in Lemma 3.37 of Stroock [116], Theorem 4.1 of Varadhan [122], and Lemma 3.2.13 of Deuschel and Stroock [42]. The proof relies on a theorem of Lusin which asserts that for any bounded measurable function $V(x)$ and any $\epsilon > 0$, there exists a continuous function $V_{\epsilon}(x)$ such that $V_{\epsilon}(x) = V(x)$ for all $x$ in some closed subset $K_{\epsilon} \subseteq \mathcal{X}$ with probability at least $1 - \epsilon$ under both $P$ and $M$. One may assume that $V_{\epsilon}(x)$ is bounded between the extremes of $V(x)$ so that $[(P, V_{\epsilon}) - \Lambda_M(V_{\epsilon})]$ converges to $[(P, V) - \Lambda_M(V)]$ as $\epsilon \to 0$. If $V(x)$ is bounded and continuous then $(P, V)$ is linear and continuous in $P$ and $\Lambda_M(V) = \log(\int e^V \, dM)$ is continuous and concave in $M$. It follows that $I(P|M) = \sup_{V} [(P, V) - \Lambda_M(V)]$ is convex and lower semicontinuous in the pair $(P, M)$. Donsker and Varadhan [46] also proved that when $\mathcal{X}$ is a complete separable metric (Polish) space, then the level sets $\{P : I_M(P) \leq L\}$ are compact for the weak topology. Compactness of the level sets is what characterizes good rate functions in large deviation theory.

E. Conditional information

Let $X$ be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{B})$. Let $P_\mathcal{G}$ denote the restriction of $P$ to some sub-$\sigma$-field $\mathcal{G}$ of $\mathcal{F}$, and let $\mu(dx|\mathcal{G})$ be a transition kernel from $(\Omega, \mathcal{G})$ to $(\mathcal{X}, \mathcal{B})$. Thus $\mu(dx|\mathcal{G})$ is a measurable function from $(\Omega, \mathcal{G})$ to the space of measures on $\mathcal{X}$, when that space is equipped with the smallest $\sigma$-field such that the evaluation map $B \mapsto Q(B)$ is measurable in $Q$. The conditional information of $X$ given $\mathcal{G}$ is defined as the maximum growth exponent

$$I_\mu(X|\mathcal{G}) = \sup_{q(x|\mathcal{G})} E_P\{\log q(X|\mathcal{G})\},$$

where the supremum is taken over all $B \otimes \mathcal{G}$-measurable functions $q(x|\mathcal{G}) \geq 0$ such that $\int q(x|\mathcal{G}) \mu(dx|\mathcal{G}) \leq 1$ almost surely. If the conditional density $p(x|\mathcal{G})$ of $P(dx|\mathcal{G})$ $P_\mathcal{G}(d\omega)$ relative to $\mu(dx|\mathcal{G})$ $P_\mathcal{G}(d\omega)$ exists, then $p(x|\mathcal{G})$ is conditionally log-optimum given $\mathcal{G}$ and

$$I_\mu(X|\mathcal{G}) = E_P\{\log p(X|\mathcal{G})\}.$$

One may interpret $I_\mu(X|\mathcal{G})$ as the maximum growth exponent of compounded wealth in independent rounds of gambling on $X$ with knowledge of $\mathcal{G}$ against odds defined by the conditional measure $\mu(dx|\mathcal{G})$. Alternatively, $I_\mu(X|\mathcal{G})$ is the information of $P(dx|\mathcal{G})$ $P_\mathcal{G}(d\omega)$ relative to the measure $\mu(dx|\mathcal{G})$ $P_\mathcal{G}(d\omega)$ on $(\mathcal{X}, \mathcal{B}) \times (\Omega, \mathcal{G})$.

If $\mu(dx|\mathcal{G})$ is normalized (or, more generally, if the total mass $\mu(\mathcal{X}|\mathcal{G}) = \int \mu(dx|\mathcal{G})$ is bounded), then $\mu(dx|\mathcal{G}) P_\mathcal{G}(d\omega)$ is a finite measure on $\mathcal{B} \otimes \mathcal{G}$ and, by Theorem 2,

$$I_\mu(X|\mathcal{G}) = \sup_{\alpha, \epsilon} I_\mu(\alpha(X)|\mathcal{E}),$$

(8)
where the supremum is taken over finite partitions \( \alpha \) with atoms in a generating subfield of \( \mathcal{B} \) and finite subfields \( \mathcal{E} \) of a generating subfield of \( \mathcal{G} \). In fact, if \( \mathcal{B} \) and \( \mathcal{G} \) are generated by countable subfields \( \mathcal{A} \) and \( \mathcal{H} \), then there exists a sequence of simple \( \mathcal{A} \otimes \mathcal{H} \)-measurable conditional density functions \( q_k(x|\mathcal{G}) \) such that for every probability measure \( P \) on \( (\Omega, \mathcal{F}) \),

\[
I_\mu(X|\mathcal{G}) = \sup_k E_P\{\log q_k(X|\mathcal{G})\}. \tag{9}
\]

In the general case when \( \mu(dx|\mathcal{G}) \) is not normalized, let \( \gamma(x|\mathcal{G}) \geq 0 \) be the conditional density function of a normalized or sub-normalized transition kernel \( M(dx|\mathcal{G}) = \gamma(x|\mathcal{G}) \mu(dx|\mathcal{G}) \) from \( (\Omega, \mathcal{G}) \) to \( (X, \mathcal{B}) \). Then \( I_\mu(X|\mathcal{G}) = E\{\log \gamma(X|\mathcal{G})\} + I_M(X|\mathcal{G}) \) and there exist nonnegative \( \mathcal{A} \otimes \mathcal{H} \)-measurable functions \( g_k(x|\mathcal{G}) \) such that \( I_M(X|\mathcal{G}) = \sup_k E\{\log g_k(X|\mathcal{G})\} \) for any probability measure \( P \) on \( (\Omega, \mathcal{F}) \) such that \( E\{\log \gamma(X|\mathcal{G})\} > -\infty \). The approximation property (9) holds if for all \( k \) we set \( q_k(x|\mathcal{G}) = \gamma(x|\mathcal{G}) g_k(x|\mathcal{G}) \).

Suppose the sub-\( \sigma \)-fields \( \mathcal{G}_k \) are monotonically increasing to the limiting \( \sigma \)-field \( \mathcal{G} = \bigvee_k \mathcal{G}_k \). Let \( \mu(dx|\mathcal{G}_K) \) be a transition kernel from \( (\Omega, \mathcal{G}_K) \) to \( (X, \mathcal{B}) \). The conditional information \( I_\mu(X|\mathcal{G}_k) \) is defined for all \( k \geq K \) as the maximum growth exponent of conditional density functions \( q(x|\mathcal{G}_k) \) such that \( f q(x|\mathcal{G}_k) \mu(dx|\mathcal{G}_K) \leq 1 \) almost surely. Since \( \bigvee_k \mathcal{G}_k \) is a generating subfield of \( \mathcal{G} \), it follows by application of (8) that when \( \mu(dx|\mathcal{G}) \) is normalized,

\[
I_\mu(X|\mathcal{G}_k) > I_\mu(X|\mathcal{G}) \quad \text{as} \quad k \to \infty. \tag{10}
\]

This no-gap property also holds in the unnormalized case, provided \( I_\mu(X|\mathcal{G}_K) > -\infty \). Indeed, let \( \gamma(x|\mathcal{G}_K) \geq 0 \) be a conditional density such that \( M(dx|\mathcal{G}_K) = \gamma(x|\mathcal{G}_K) \mu(dx|\mathcal{G}_K) \) is a subnormalized transition kernel from \( (\Omega, \mathcal{G}_K) \) to \( (X, \mathcal{B}) \) and \( E\{\log \gamma(X|\mathcal{G}_K)\} > -\infty \). Then \( I_\mu(X|\mathcal{G}_k) = E\{\log \gamma(X|\mathcal{G}_K)\} + I_M(X|\mathcal{G}_K) \) and \( I_M(X|\mathcal{G}_K) \nless I_M(X|\mathcal{G}) \), hence \( I_\mu(X|\mathcal{G}_k) \nless I_\mu(X|\mathcal{G}) \). Note that we have proved the no-gap property (10) without recourse to martingale theory or the combinatorial argument of Dobrushin [44] which appears at the end of Chapter 2 of Pinsker [101], in the paper of Csiszár [31], and in Lemma 5.2.2 of Gray [59].

**F. Convergence in information**

Let \( P \) and \( Q \) be probability distributions and let \( m \) be a \( \sigma \)-finite reference measure on the space \( \mathcal{X} \). If \( I_m(Q) \) is well defined and finite then the function \( V_Q(x) = \log[dQ/dm(x)] \) may be regarded as the gradient and the divergence functional \( I_Q(P) \) may be regarded as the second order remainder term in the Taylor series expansion of \( I_m(P) \) near \( Q \) since

\[
I_m(P) = I_m(Q) + (P - Q, V_Q) + I_Q(P) = (P, V_Q) + I_Q(P),
\]

where \( (P, V) = \int V \ dP \). This expansion is equivalent to the formal identity

\[
\int dP \log \left( \frac{dP}{dm} \right) = \int dQ \log \left( \frac{dQ}{dm} \right) + \int (dP - dQ) \log \left( \frac{dQ}{dm} \right) + \int dP \log \left( \frac{dP}{dQ} \right).
\]

Since \( I_Q(P) \geq 0 \) with equality iff \( P = Q \), we see that \( I_Q(P) \) is the vertical distance at \( P \) between the graph of the convex functional \( I_m(P) \) and the hyperplane with slope \( V_Q \) that is supporting the graph at \( Q \). The contact point \( Q \) is unique since the functional \( I_Q(P) \) is strictly convex in \( P \).

The divergence \( I(P|Q) = I_Q(P) \) is bounded below by a quadratic function of the variation distance \( \|P - Q\| \). The variation distance is defined as

\[
\|P - Q\| = \int |p - q| \ dm = |P(A) - Q(A)| + |P(A^c) - Q(A^c)|
\]
where \( p = dP/dm \) and \( q = dQ/dm \) are the densities of \( P \) and \( Q \) relative to a dominating \( \sigma \)-finite measure \( m \) and where \( A^c \) is the complement of \( A = \{ p < q \} \). Csiszár [30], Kemperman [71] and Kullback [77] proved that

\[
I(P|Q) \geq \frac{1}{2} \|P - Q\|^2
\]

when \( P \) and \( Q \) are probability measures, and that the coefficient \( 1/2 \) is best possible. This result remains valid if \( Q \) is not normalized but a subprobability measure.

**Lemma 2.** If \( P \) is a probability and \( Q \) is a subprobability measure on \( \mathcal{X} \) then

\[
\|P - Q\| \leq \sqrt{2I_Q(P)}.
\]

**Proof:** Following Kemperman [71], we observe that \( 3(f - 1)^2 \leq (4 + 2f)(f \log f - f + 1) \) if \( f \geq 0 \). Setting \( f = p/q \) and multiplying by \( q^2 \) yields

\[
3(p - q)^2 \leq [4q + 2p][p \log(p/q) - p + q].
\]

We may assume without loss of generality that \( m \) is normalized so that by the Cauchy-Schwartz inequality,

\[
\int |p - q| dm \leq \left[ \frac{4Q(\mathcal{X}) + 2P(\mathcal{X})}{3} \right]^{\frac{1}{2}} [I_Q(P) - P(\mathcal{X}) + Q(\mathcal{X})]^{\frac{1}{2}}.
\]

The lemma follows since \( 0 \leq Q(\mathcal{X}) \leq 1 = P(\mathcal{X}) \). □

Note that the inequality \( \|P - Q\| \leq \sqrt{2I_Q(P)} \) is still true if \( I_Q(P) = \int p \log(p/q) dm \) is replaced by the smaller functional \( \int [p \log(p/q) - p + q] dm = I_Q(P) - P(\mathcal{X}) + Q(\mathcal{X}) \).

The following result was originally proved for probability distributions \( P, Q \) by Pinsker [101] and in enhanced form by Barron [9]. We now assume \( Q \) is a subprobability measure.

**Lemma 3.** If the distribution \( P \) is dominated by a subprobability measure \( Q \) on \( \mathcal{X} \), then

\[
I_Q(P) \leq E_P \left\{ \left| \log \left( \frac{dP}{dQ} \right) \right| \right\} \leq I_Q(P) + \sqrt{2I_Q(P)}.
\]

**Proof:** The left inequality in (11) is obvious in view of (3). To prove the right inequality, let \( A = \{ dP/dQ < 1 \} \) and observe that

\[
\|P - Q\| = |P(A) - Q(A)| + |P(A^c) - Q(A^c)| \geq 2[Q(A) - P(A)]
\]

since \( |P(A^c) - Q(A^c)| = 1 - P(A) + Q(A) - Q(\mathcal{X}) \geq Q(A) - P(A) \geq 0 \). Furthermore

\[
\int_A dP \log \left( \frac{dQ}{dP} \right) \leq \int_A dP \left( \frac{dQ}{dP} - 1 \right) = Q(A) - P(A),
\]

and consequently

\[
\int dP \left| \log \left( \frac{dP}{dQ} \right) \right| = \int dP \log \left( \frac{dP}{dQ} \right) + 2 \int_A dP \log \left( \frac{dQ}{dP} \right) \leq I_Q(P) + 2[Q(A) - P(A)].
\]

18
Now $2[Q(A) - P(A)] \leq \|P - Q\|$ and $\|P - Q\| \leq \sqrt{2I_Q(P)}$ by Lemma 2.

We say that a sequence of subprobability measures $Q_k$ converges in information to $P$ if

$$I(P|Q_k) \to 0 \quad \text{as} \quad k \to \infty.$$ 

By Lemma 2, this happens iff $P$ is dominated by $Q_k$ eventually for large $k$ and

$$\log \left( \frac{dP}{dQ_k} \right) \to 0 \quad \text{in} \quad L^1(P) \quad \text{as} \quad k \to \infty.$$ 

Since $\|P - Q_k\| \leq \sqrt{2I(P|Q_k)}$, convergence $Q_k \to P$ in information implies convergence in variation norm. The total mass $Q_k(\mathcal{X})$ will converge to 1 and the normalized measure $dM_k = dQ_k/Q_k(\mathcal{X})$ will converge in variation norm to $P$ since the nonnegative quantities $-\log Q_k(\mathcal{X})$ and $I(P|M_k) \geq \|P - M_k\|^2/2$ converge to 0:

$$0 \leq -\log Q_k(\mathcal{X}) + I_M(P) = I_{Q_k}(P) \to 0.$$ 

It should perhaps be emphasized that our definition of convergence in information differs from that of Csiszár [35]. According to [35], a sequence of probability measures $P_k$ converges in information to a probability measure $Q$ if

$$I(P_k|Q) \to 0 \quad \text{as} \quad k \to \infty.$$
III. LOG-OPTIMUM SELECTIONS IN CONVEX FAMILIES

A. Karush-Kuhn-Tucker conditions for log-optimality

Let $Q$ be a family of nonnegative random variables defined on the probability space $(\Omega, \mathcal{F}, P)$. An element $q^* \in Q$ is called log-optimum in $Q$ if $q^* > 0$ almost surely and no alternative selection in $Q$ can improve upon $q^*$ in growth exponent:

$$E \left\{ \log \left( \frac{q}{q^*} \right) \right\} \leq 0, \quad \text{for all } q \in Q.$$

If the expectation $E\{\log q\}$ is well defined for all $q \in Q$, then any log-optimum $q^* \in Q$ attains the maximum growth exponent

$$W_Q(P) = \sup_{q \in Q} E_P\{\log q\}.$$

Conversely, if $W_Q(P)$ is well defined and finite then any $q^* \in Q$ which attains the maximum is log-optimum in $Q$. However, a log-optimum $q^*$ may exist even if the expectation $E\{\log q\}$ is not well defined for all $q \in Q$.

A sufficient condition for log-optimality of $q^*$ is that $E\{q/q^*\} \leq 1$ for all other $q \in Q$. This follows from Jensen’s inequality

$$E \left\{ \log \left( \frac{q}{q^*} \right) \right\} \leq \log \left( E \left\{ \frac{q}{q^*} \right\} \right) \leq \log 1 = 0. \quad (12)$$

Bell and Cover [12] proved that these sufficient conditions are also necessary when the family $Q$ is convex. In fact, they are the Karush-Kuhn-Tucker (KKT) conditions for elements $q^*$ that maximize the concave functional $W(q) = E\{\log q\}$ over the convex set $Q$. Previously, Bell and Cover [11] and Breiman [20] had already characterized log-optimum portfolios in investment by means of such conditions. See also Chapter II of Algoet [2]. Whittle [128] derived the KKT conditions in the context of optimal experimental design and Lindsay [86] applied Whittle's results to maximum likelihood estimation.

**Theorem 6.** Let $Q$ be a convex family of nonnegative random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $q^*$ be a random variable in $Q$ such that $q^* > 0$ almost surely. Then $q^*$ is log-optimum in $Q$ iff $q^*$ satisfies the Karush-Kuhn-Tucker (KKT) conditions

$$E \left\{ \frac{q}{q^*} \right\} \leq 1, \quad \text{for all } q \in Q. \quad (13)$$

**Proof:** The functional $W(q) = E\{\log q\}$ is concave on the convex set $Q$. The directional derivative of $W$ as we move away from $q_0$ along the segment $q_\epsilon = (1 - \epsilon)q_0 + \epsilon q_1$ toward some alternative $q_1$ in $Q$ is defined as

$$D(q_1|q_0) = \lim_{\epsilon \to 0} \frac{W(q_\epsilon) - W(q_0)}{\epsilon}.$$

The limit exists and exceeds $W(q_1) - W(q_0)$ since the difference quotient $\epsilon^{-1}[W(q_\epsilon) - W(q_0)]$ of the concave function $W(q_\epsilon)$ is monotonically increasing from $W(q_1) - W(q_0)$ to $D(q_1|q_0)$ as $\epsilon$ decreases from $\epsilon = 1$ to $\epsilon = 0$. The KKT conditions assert that $q^*$ is log-optimum in
\( \mathcal{Q} \) iff \( D(q|q^*) \leq 0 \) for all \( q \in \mathcal{Q} \). Since also \( D(q^*|q) \geq W(q^*) - W(q) \geq 0 \) for all \( q \in \mathcal{Q} \), the log-optimum \( q^* \) will be a saddle point of \( D(q_1|q_0) \): for all \( q_0, q_1 \in \mathcal{Q} \) we have

\[
D(q_1|q^*) \leq 0 = D(q^*|q^*) = 0 \leq D(q^*|q_0).
\]

One can express \( D(q|q^*) \) in terms of the relative change \( \Delta = (q - q^*)/q^* \) as

\[
D(q|q^*) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} E\{\log(1 + \epsilon \Delta)\}.
\]

The limit of the expectations is equal to the expectation of the limit. Indeed, if \( a > 0 \) then by a Taylor series expansion as in Theorem 1 of [3]

\[
\epsilon \Delta \geq \log(1 + \epsilon \Delta) \geq \log(1 + \epsilon (\Delta \wedge a)) = \epsilon (\Delta \wedge a) - \frac{1}{2} \theta \epsilon^2 (\Delta \wedge a)^2 \quad \text{(for some } 0 < \theta < 1) \geq \epsilon (\Delta \wedge a) - \frac{1}{2} \epsilon^2 a^2.
\]

Choosing \( a = a(\epsilon) \) so that \( a(\epsilon) \uparrow \infty \) and \( \epsilon a(\epsilon)^2 \to 0 \) as \( \epsilon \downarrow 0 \) proves that \( D(q|q^*) = E\{\Delta\} \).

The KKT conditions \( D(q|q^*) \leq 0 \) or \( E\{\Delta\} \leq 0 \) are equivalent to (13).

The theorem also follows from the following argument which appears in Theorem 3 of Breiman [20]. Observe that

\[
\frac{q}{q^*} = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \log \left( 1 + \frac{\epsilon q}{(1 - \epsilon)q^*} \right) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \log \left( \frac{q}{q^*} \right) + \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \log \left( \frac{1}{1 - \epsilon} \right),
\]

and consequently, by Fatou’s lemma,

\[
E \left\{ \frac{q}{q^*} \right\} \leq \liminf_{\epsilon \downarrow 0} \frac{1}{\epsilon} E \left\{ \log \left( \frac{q}{q^*} \right) \right\} + 1.
\]

Since \( E\{\log(q_e/q^*)\} \leq 0 \), we may conclude that \( E\{q/q^*\} \leq 1 \).

If the line segment from \( q \) to \( q^* \) can be extended beyond \( q^* \) in \( \mathcal{Q} \), then \( E\{q/q^*\} = 1 \) since the two-sided derivative \( E\{\Delta\} \) of \( E\{\log(q_e/q^*)\} \) must vanish at \( \epsilon = 0 \). Thus the KKT condition (13) holds with equality for all \( q \in \mathcal{Q} \) if the log-optimum \( q^* \) falls in the algebraic interior of \( \mathcal{Q} \).

Note that \( q^* \) is log-optimum in \( \mathcal{Q} \) iff \( q^* \) is log-optimum in the family \( \mathcal{Q}' \) of random variables \( q' \) such that \( 0 \leq q' \leq q \) for some \( q \in \mathcal{Q} \). The family \( \mathcal{Q}' \) is down-monotone: if \( 0 \leq q' \leq q \) and \( q \in \mathcal{Q}' \) then also \( q' \in \mathcal{Q}' \). It is often convenient to replace \( \mathcal{Q} \) by its down-monotone hull \( \mathcal{Q}' \) so that \( \mathcal{Q} \) itself is down-monotone.

**B. Existence and uniqueness of log-optimum selections**

Let \( \mathcal{Q} \) be a convex family of nonnegative random variables on the probability space \((\Omega, \mathcal{F}, P)\). The log-optimum \( q^* \) in \( \mathcal{Q} \) is unique up to almost sure equivalence, if it exists. Indeed, if Jensen’s inequality (12) holds with equality then \( q/q^* = 1 \) almost surely since \( \log(\cdot) \) is a strictly concave function. This was already observed by Breiman [20].

A log-optimum \( q^* \) may fail to exist in \( \mathcal{Q} \). However, if \( W_\mathcal{Q}(P) \) is well defined and finite then a log-optimum \( q^* \) exists in a certain closure \( \overline{\mathcal{Q}} \) of \( \mathcal{Q} \). Indeed, let \( \{q_k\} \) be a sequence in
Q such that \( E \{ \log q_k \} \) converges to the maximum growth exponent \( W_Q(P) \). By Lemma 1 of Gillespie [56], \( \log q_k \) is a Cauchy sequence in measure and some subsequence converges almost surely. The limit \( \log q^* \) is obviously independent of the particular choice of elements \( q_k \) such that \( E \{ \log q_k \} \to W_Q(P) \). Theorem 1 of Kieffer [72] implies that \( \log q_k \) is actually a Cauchy sequence in \( L^1(P) \). Thus the limit of the expectations \( W_Q(P) = \lim_k E \{ \log q_k \} \) is equal to the expectation of the limit \( E \{ \log q^* \} \). It follows that \( q^* \) is log-optimum in the family \( \tilde{Q} \) of random variables \( \tilde{q} \geq 0 \) such that \( \log \tilde{q} \) belongs to the closure of the family \( \{ \log q : q \in Q \} \) in \( L^1(P) \).

**Theorem 7.** Let \( Q \) be a convex family of nonnegative random variables on \( (\Omega, \mathcal{F}, P) \) and let \( \tilde{Q} \) denote the family of random variables \( \tilde{q} > 0 \) such that \( \log(q_k/\tilde{q}) \to 0 \) in \( L^1(P) \) for some sequence \( \{ q_k \} \) in \( Q \).

(a). The family \( \tilde{Q} \) is convex, \( \overline{Q} \subseteq \tilde{Q} \), and \( W_{\tilde{Q}}(P) = W_Q(P) \).

(b). If \( W_Q(P) \) is well defined and finite then there exists a log-optimum \( q^* \) in \( \tilde{Q} \). The random variable \( q^* \) is unique up to almost sure equivalence, and for any \( q_k \in Q \) such that \( E \{ \log q_k \} \) converges to the maximum growth exponent \( W_{\tilde{Q}}(P) \) we have mean convergence

\[
\log q_k \to \log q^* \text{ in } L^1(P).
\]

**Proof of (a):** It is obvious that \( \overline{Q} \subseteq \tilde{Q} \) and \( W_{\tilde{Q}}(P) = W_Q(P) \). To prove that \( \tilde{Q} \) is convex, consider random variables \( \tilde{q} \) and \( \tilde{q}' \) in \( \tilde{Q} \) and sequences \( \{ q_k \} \) and \( \{ q'_k \} \) in \( Q \) such that

\[
\log(q_k/\tilde{q}) \to 0 \text{ in } L^1(P), \quad \log(q'_k/\tilde{q}') \to 0 \text{ in } L^1(P).
\]

We may assume almost sure convergence as well, since we may pass to subsequences. If \( \lambda, \lambda' \) are nonnegative weights with sum \( \lambda + \lambda' = 1 \), then \( (\lambda q_k + \lambda' q'_k)/(\lambda \tilde{q} + \lambda' \tilde{q}') \) is sandwiched between \( q_k/\tilde{q} \) and \( q'_k/\tilde{q}' \). Both sequences \( \log(q_k/\tilde{q}) \) and \( \log(q'_k/\tilde{q}') \) vanish pointwise and in mean as \( k \to \infty \), so they are uniformly integrable and

\[
\log \left( \frac{\lambda q_k + \lambda' q'_k}{\lambda \tilde{q} + \lambda' \tilde{q}'} \right) \to 0 \quad \text{almost surely and in } L^1(P).
\]

Thus \((\lambda \tilde{q} + \lambda' \tilde{q}') \) belongs to \( \tilde{Q} \), proving that \( \tilde{Q} \) is convex.

**Proof of (b):** Select \( q_k \in Q \) such that \( E \{ \log q_k \} \) converges to the maximum growth exponent \( W_{\tilde{Q}}(P) \) as \( k \to \infty \). By the dominated convergence theorem (or by Theorem 2 of [3]), there exists some \( q^*_k \in Q \) such that \( q^*_k \) is log-optimum in the convex hull of \( \{ q_1, \ldots, q_k \} \). If \( k \leq l \) then \( dQ_{k,l} = (q^*_k/q^*_l) dP \) is a subprobability since \( E \{ q^*_k/q^*_l \} \leq 1 \). If \( k \leq l \) and \( \inf(k,l) \to \infty \) then \( Q_{k,l} \) converges in information to \( P \) since

\[
I_{Q_{k,l}}(P) = E_P \left\{ \log \left( \frac{q^*_k}{q^*_l} \right) \right\} = W_{q^*_k}(P) - W_{q^*_l}(P) \to 0.
\]

By Lemma 3, if \( k \leq l \) then

\[
I(P|Q_{k,l}) \leq E_P \left| \log \left( \frac{dP}{dQ_{k,l}} \right) \right| \leq I(P|Q_{k,l}) + \sqrt{2I(P|Q_{k,l})}.
\]

(14)

The sequence \( \{ q^*_k \} \) is a Cauchy sequence in \( L^1(P) \) because

\[
\log \left( \frac{q^*_k}{q^*_l} \right) = \log \left( \frac{dP}{dQ_{k,l}} \right) \to 0 \quad \text{in } L^1(P)
\]
when \( k \leq l \) and \( \inf(k, l) \to \infty \). Thus there exists a random variable \( q^* \geq 0 \) such that
\[
\log q_k \to \log q^* \quad \text{in } L^1(P).
\]
The expectation of the limit \( \log q^* \) is equal to the limit of the expectations, hence \( E_P(\log q^*) = W_Q(P) \). Since \( q^* \in \bar{Q} \) attains the maximum growth exponent \( W_Q(P) = W_{\bar{Q}}(P) \), we see that \( q^* \) is log-optimum in \( \bar{Q} \) and hence, by Theorem 6, that \( E\{q/q^*\} \leq 1 \) for all \( q \in \bar{Q} \). The subprobabilities \( dQ_k = (q_k/q^*) dP \) converge in information to the true distribution \( dP \) and hence \( \log(q_k/q^*) \to 0 \) in \( L^1(P) \). This completes the proof of the theorem.

Note that \( dQ_k = (q_k/q^*) dP \) also converges to \( dP \) in variation norm. Thus
\[
E_P \left| \frac{dQ_k}{dP} - 1 \right| = E_P \left| \frac{q_k}{q^*} - 1 \right| \to 0 \quad \text{as } k \to \infty.
\]
Thus the KKT conditions are asymptotically satisfied with equality:
\[
E_P \left\{ \frac{dQ_k}{dP} \right\} = E_P \left\{ \frac{q_k}{q^*} \right\} \to 1.
\]

Kieffer [72] did not explicitly mention convex families, but formulated his results instead for "log-convex" families of random variables. A family \( \mathcal{L} \) is called log-convex if the family of random variables \( \{e^{-\ell} : \ell \in \mathcal{L}\} \) is convex, or equivalently if there exists a convex family \( \mathcal{Q} \) of nonnegative random variables such that
\[
\mathcal{L} = \{\ell : \ell = -\log q \text{ for some } q \in \mathcal{Q}\}.
\]
The following example illustrates the correspondence between convex and log-convex families: If \( X \) is a random variable with values in a finite set \( \mathcal{X} \) and if \( q(x) \) ranges over all the probability mass functions on \( \mathcal{X} \), then the likelihood \( q(X) \) ranges over a convex family of nonnegative random variables and the negative log-likelihood or ideal codeword length \( -\log q(X) \) ranges over a log-convex family. Kieffer proved that if a log-convex family \( \mathcal{L} \) is closed in \( L^1 \) and \( \inf_{\ell \in \mathcal{L}} E\{\ell\} \) is well defined and finite then the minimum expectation is attained by a unique random variable \( \ell^* \in \mathcal{L} \). In fact, \( \ell_k \to \ell^* \) in \( L^1 \) for any \( \ell_k \in \mathcal{L} \) such that \( E\{\ell_k\} \) decreases to the minimum expectation. The assumption that \( \mathcal{L} \) is closed in \( L^1 \) can be removed since the \( L^1 \)-closure of a log-convex family is log-convex.

C. Lower semicontinuity of \( W_Q(P) \)

The maximum growth exponent \( W_Q(P) = \sup_{q \in \mathcal{Q}} E_P\{\log q(X)\} \) is convex in \( P \) since it is the supremum of affine functionals \( E_P\{\log q(X)\} \). We shall prove that \( W_Q(P) \) is bounded below and lower semicontinuous in \( P \) if \( \mathcal{Q} \) is convex and some function \( \gamma(x) \) in \( \mathcal{Q} \) is bounded below by a positive constant. First, we generalize Theorem 2 and Theorem 3.

**Theorem 8.** Let \( \mathcal{Q} \) be a convex down-monotone family of nonnegative measurable functions on a measurable space \((\mathcal{X}, \mathcal{B})\), and let \( \mathcal{A} \) be a generating subfield of \( \mathcal{B} \). If some function \( \gamma(x) \) in \( \mathcal{Q} \) is bounded below by a positive constant, then the supremum \( W_Q(P) = \sup_{q \in \mathcal{Q}} E_P\{\log q(X)\} \) may be taken over simple \( \mathcal{A} \)-measurable functions in \( \mathcal{Q} \) with rational values. If the generating subfield \( \mathcal{A} \) is countable, then there exists a sequence \( \{q_k(x)\}_{k \geq 1} \) of simple \( \mathcal{A} \)-measurable functions in \( \mathcal{Q} \) such that for every probability measure \( P \) on \( \mathcal{X} \),
\[
W_Q(P) = \sup_k E_P\{\log q_k(X)\}.
\]
If \( \gamma(x) \) is an arbitrary function in \( \mathcal{Q} \), then there exists a sequence of functions \( q_k(x) \) in \( \mathcal{Q} \) such that (15) holds for every probability distribution \( P \) such that \( E_P\{\log \gamma(X)\} > -\infty \).

Proof: We generalize the proof of Theorem 2 and Theorem 3, which deal with the special case when \( \mathcal{Q} \) is the family \( \mathcal{Q}(m) \) and \( W_{\mathcal{Q}}(P) \) is the information functional \( I_m(P) \).

Suppose \( \log \gamma(x) \) is bounded below. For each finite partition \( \alpha \) of \( \mathcal{X} \) let \( \mathcal{Q}_\alpha \) denote the family of functions in \( \mathcal{Q} \) that take on a constant rational value on each atom of \( \alpha \). Given a probability distribution \( P \) on \( \mathcal{X} \), let \( q^*_\alpha(x) \) be log-optimum in the convex closure of the countable family \( \mathcal{Q}_\alpha \). Reasoning as in Theorem 2, we see that

\[
W_{\mathcal{Q}}(P) = \sup_{\alpha \in \Pi(A)} W_{\mathcal{Q}_\alpha}(P) = \sup_{\alpha \in \Pi(A)} E_P\{\log q^*_\alpha(X)\}.
\]

If the generating field \( \mathcal{A} \) is countable, then \( W_{\mathcal{Q}}(P) = \sup_{\alpha \in \Pi(A)} W_{\mathcal{Q}_\alpha}(P) \) is the supremum over a countable family of countable families. Enumerating the union, we obtain a sequence \( \{q_k(x)\}_{k \geq 1} \) of simple \( \mathcal{A} \)-measurable functions in \( \mathcal{Q} \) such that (15) holds. In fact, if \( \{\alpha_k\}_{k \geq 1} \) is a sequence of finer and finer partitions that asymptotically generate \( \mathcal{B} \) and if \( P \) is an arbitrary probability measure on \( \mathcal{X} \), then

\[
\log q^*_{\alpha_k}(X) \rightarrow \log q^*(X) \quad \text{in} \quad L^1(P),
\]

and each \( \log q^*_\alpha(X) \) is equal to \( \lim_{j \to \infty} \log q_{k,j}(X) \) in \( L^1(P) \) for some \( q_{k,j} \in \mathcal{Q}_{\alpha_k} \).

If \( \log \gamma(x) \) is not bounded below but \( E_P\{\log \gamma(X)\} > -\infty \) then we write

\[
W_{\mathcal{Q}}(P) = E_P\{\log \gamma(X)\} + W_{\mathcal{Q}/\gamma}(P),
\]

where \( \mathcal{Q}/\gamma \) is the family of ratios \( q(x)/\gamma(x) \). The constant 1 belongs to \( \mathcal{Q}/\gamma \), and the first part of the theorem implies the existence of simple \( \mathcal{A} \)-measurable functions \( q_k(x)/\gamma(x) \) in \( \mathcal{Q}/\gamma \) such that \( W_{\mathcal{Q}/\gamma}(P) = \sup_k E_P\{\log[q_k(X)/\gamma(X)]\} \}. The theorem follows.

The space \( \mathcal{P} \) of probability measures on \( \mathcal{X} \) will be equipped with either the strong or the weak topology. By strong convergence \( P_k \rightarrow P \) is meant that \( P_k(B) \rightarrow P(B) \) for every measurable set \( B \subseteq \mathcal{X} \), or equivalently that \( (P_k, V) \rightarrow (P, V) \) for every bounded measurable function \( V(x) \) on \( \mathcal{X} \). If \( \mathcal{X} \) is a metric space with its Borel \( \sigma \)-field, then by weak convergence \( P_k \rightarrow P \) is meant that \( (P_k, V) \rightarrow (P, V) \) for every bounded continuous function \( V(x) \), or equivalently that \( P_k(B) \rightarrow P(B) \) for every Borel set \( B \) whose boundary \( \partial B \) has vanishing probability \( P(\partial B) = 0 \).

A functional \( W(P) \) is called lower semicontinuous if \( \liminf P_k \geq W(P) \) whenever \( P_k \rightarrow P \) in \( \mathcal{P} \). It is equivalent to require that the level sets \( \{ P : W(P) \leq L \} \) are closed in \( \mathcal{P} \), or that the epigraph \( \{(P, r) : W(P) \leq r \} \) is closed in \( \mathcal{P} \times \mathbb{R} \). The supremum of any family of continuous or lower semicontinuous functions is lower semicontinuous.

Kullback [76] proved that the divergence \( I(P|M) \) between probability measures \( P \) and \( M \) is lower semicontinuous in the pair \( (P, M) \) when \( \mathcal{P} \) is equipped with the strong topology. Formally, if \( P_k \rightarrow P \) and \( M_k \rightarrow M \) in the strong sense, then

\[
\liminf_k I(P_k|M_k) \geq I(P|M).
\]

Csiszár [33] and Posner [103] proved the stronger result that \( I(P|M) \) is lower semicontinuous for the weak topology. See also Gelfand and Yaglom [54], Pinsker [101], Dobrushin [45].

Theorem 9. (Lower semicontinuity of \( W_{\mathcal{Q}}(P) \))

24
Let $\mathcal{Q}$ be a convex family of nonnegative measurable functions on a measurable space $\mathcal{X}$. If some $\gamma(x)$ in $\mathcal{Q}$ is bounded below by a positive constant, then the maximum growth exponent $W_{\mathcal{Q}}(P)$ is convex, bounded below, and lower semicontinuous in $P$ when the space of probability measures on $\mathcal{X}$ is equipped with the strong topology. If $\mathcal{X}$ is a metric space with its Borel $\sigma$-field, then $W_{\mathcal{Q}}(P)$ is lower semicontinuous in $P$ even for the weak topology.

Proof: The supremum $W_{\mathcal{Q}}(P) = \sup_{q \in \mathcal{Q}} E_P\{\log q(X)\}$ may be taken over functions in $\mathcal{Q}$ such that $\log q(x)$ is simple and hence bounded. If $\log q(x)$ is bounded, then the affine functional $E_P\{\log q(X)\}$ is continuous in $P$ for the strong topology. The functional $W_{\mathcal{Q}}(P)$ is convex, bounded below and lower semicontinuous for the strong topology since it is the supremum of affine functionals $E_P\{\log q(X)\}$ that are bounded and continuous in $P$.

Suppose $\mathcal{X}$ is a metric space with Borel $\sigma$-field $B$. We claim that $\liminf_k W_{\mathcal{Q}}(P_k) \geq W_{\mathcal{Q}}(P)$ if $P_k$ converges weakly to $P$. Indeed, let $A_P = \{B \in B : P(\partial B) = 0\}$ denote the field of $P$-continuity sets. Then $A_P$ is a generating subfield of $B$ (see Theorem 1 of Posner [103] for a careful proof). By Theorem 8, the maximum growth exponent $W_{\mathcal{Q}}(P)$ may be taken over simple $A_P$-measurable functions in $\mathcal{Q}$. If $\log q(x)$ is a simple $A_P$-measurable function in $\mathcal{Q}$, then $E_{P_k}\{\log q(X)\} \rightarrow E_P\{\log q(X)\}$ and hence, since $W_{\mathcal{Q}}(P_k) \geq E_{P_k}\{\log q(X)\}$,

$$\liminf_k W_{\mathcal{Q}}(P_k) \geq \liminf_k E_{P_k}\{\log q(X)\} = E_P\{\log q(X)\}.$$ 

The stated claim follows by taking the supremum over all simple $A_P$-measurable functions $q(x)$ in $\mathcal{Q}$. The theorem follows since $P$ was an arbitrary distribution on $\mathcal{X}$. 

### D. Conditional log-optimality

Let $\mathcal{Q}$ be a family of nonnegative random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G}$ be a sub-$\sigma$-field of $\mathcal{F}$. A selection $q^* \in \mathcal{Q}$ is said to be conditionally log-optimum in $\mathcal{Q}$ given the side information $\mathcal{G}$ if $q^* > 0$ almost surely and

$$E\left\{\log\left(\frac{q}{q^*}\right)\bigg|\mathcal{G}\right\} \leq 0, \quad \text{for all } q \in \mathcal{Q}.\]$$

We say that $\mathcal{Q}$ is $\mathcal{G}$-convex if $\mathcal{Q}$ is closed under taking convex combinations with $\mathcal{G}$-measurable weights: If $q, q' \in \mathcal{Q}$ and if $\lambda, \lambda'$ are nonnegative $\mathcal{G}$-measurable random variables with sum $\lambda + \lambda' = 1$, then the $\mathcal{G}$-convex combination $\lambda q + \lambda' q'$ also belongs to $\mathcal{Q}$.

**Theorem 10.** If the family $\mathcal{Q}$ is $\mathcal{G}$-convex then so is its closure $\overline{\mathcal{Q}}$. Any log-optimum $q^*$ in $\overline{\mathcal{Q}}$ is conditionally log-optimum given $\mathcal{G}$ and satisfies the conditional KKT conditions

$$E\left\{\frac{q}{q^*}\bigg|\mathcal{G}\right\} \leq 1 \quad \text{almost surely, for all } q \in \mathcal{Q}. \quad (16)$$

**Proof:** If $\mathcal{Q}$ is $\mathcal{G}$-convex then $\overline{\mathcal{Q}}$ is likewise since one may consider convex combinations with random $\mathcal{G}$-measurable rather than constant weights in the proof of Theorem 7(a).

Suppose $q^*$ is log-optimum in $\overline{\mathcal{Q}}$ and there exists some alternative $q \in \overline{\mathcal{Q}}$ such that the event $B$ where $E\{q/q^* | \mathcal{G}\} > 1$ has positive probability. Let $1_B$ and $1_{B^c}$ denote the indicator functions of the $\mathcal{G}$-measurable event $B$ and its complement $B^c$, and for $0 < \epsilon < 1$ let

$$q_\epsilon = [(1 - \epsilon)q^* + \epsilon q]1_B + q^*1_{B^c} = q^* + \epsilon(q - q^*)1_B.$$
Clearly \( q_\epsilon \in \mathcal{Q} \) since \( \mathcal{Q} \) is \( \mathcal{G} \)-convex. Reasoning as in the proof of Theorem 6, one obtains
\[
\frac{d}{d\epsilon} E \left\{ \log \left( \frac{q_\epsilon}{q^*} \right) \right\}_{\epsilon=0^+} = E \left\{ \frac{(q - q^*) 1_B}{q^*} \right\} = E \left\{ E \left\{ \frac{q - q^*}{q^*} \mid \mathcal{G} \right\} 1_B \right\} > 0.
\]
Thus \( E \left\{ \log(q_\epsilon/q^*) \right\} \) is positive when \( \epsilon \) is small, and this contradicts the log-optimality of \( q^* \) in \( \bar{\mathcal{Q}} \). We may conclude that the conditional KKT conditions (16) hold.

The theorem remains valid if the assumption that \( \mathcal{Q} \) is \( \mathcal{G} \)-convex is slightly weakened. It suffices to assume that \( \mathcal{Q} \) contains all combinations \( \lambda q + \lambda' q' \) where \( q, q' \) are elements of \( \mathcal{Q} \) and \( \lambda, \lambda' \) are indicator functions of complementary subsets in a generating subfield of \( \mathcal{G} \).

The maximum growth exponent can only increase when it is taken over a larger family. Let \( \{ \mathcal{Q}_k \} \) be an increasing sequence of convex families of nonnegative random variables. If \( \mathcal{Q}_k \) increases to \( \mathcal{Q} = \bigcup_k \mathcal{Q}_k \) and \(-\infty < W_{\mathcal{Q}_K}(P) \) for some \( K \), then
\[
W_{\mathcal{Q}_k}(P) / W_{\mathcal{Q}}(P) \quad \text{as} \quad k \to \infty.
\]
Suppose \( W_{\mathcal{Q}_K}(P) \) and \( W_{\mathcal{Q}}(P) \) are finite, and let \( q^*_k \) and \( q^* \) be log-optimum in the closures \( \bar{\mathcal{Q}}_k \) and \( \bar{\mathcal{Q}} \) of the families \( \mathcal{Q}_k \) and \( \mathcal{Q} \). Then, by Theorem 7,
\[
\log q^*_k \to \log q^* \quad \text{in } L^1(P).
\]
Now suppose \( \mathcal{Q}_k \) is \( \mathcal{G}_k \)-convex and the information fields \( \mathcal{G}_k \) are monotonically increasing to a limiting \( \sigma \)-field \( \mathcal{G} \). Then the log-optimum \( q^* \) in \( \mathcal{Q} \) is conditionally log-optimum given \( \mathcal{G} \) because \( \mathcal{Q} \) is closed under convex combinations with weights that are indicator functions of complementary subsets in the generating subfield \( \bigcup_k \mathcal{G}_k \) of \( \mathcal{G} = \bigvee_k \mathcal{G}_k \). In fact, \( \mathcal{Q} \) is \( \mathcal{G}_k \)-convex for all \( k \geq 0 \) and hence \( E \{ q/q^* \mid \mathcal{G}_k \} \leq 1 \) almost surely for all \( q \in \mathcal{Q} \), by Theorem 10. Since \( E \{ q/q^* \mid \mathcal{G}_k \} \) converges to \( E \{ q/q^* \mid \mathcal{G} \} \) by the martingale convergence theorem, we may conclude that \( q^* \) satisfies the conditional KKT conditions (16). We also claim that \( w_k^* = E \{ \log q^*_k \mid \mathcal{G}_k \} \) is a submartingale with respect to the \( \sigma \)-fields \( \mathcal{G}_k \). Indeed, if \( k \leq \ell \) then \( E \{ \log(q_k^*/q_\ell^*) \mid \mathcal{G}_\ell \} \leq 0 \) by Theorem 10, and consequently \( w_k^* \leq E \{ w_\ell^* \mid \mathcal{G}_k \} \) almost surely. The submartingale \( w_k^* = E \{ \log q_k^* \mid \mathcal{G}_k \} \) is uniformly integrable and converges almost surely and in \( L^1 \) to \( w^* = E \{ \log q^* \mid \mathcal{G} \} \) by Proposition IV.5.24 of Neveu [96].

**E. The no-gap property**

Let \( \mathcal{Q} \) be a convex family of nonnegative measurable functions on a measurable space \( (\mathcal{X}, \mathcal{B}) \) and let \( \{ \mathcal{G}_k \} \) be an increasing sequence of information fields with limiting \( \sigma \)-field \( \mathcal{G}_\infty \). The maximum growth exponent of measurable selections in \( \mathcal{Q} \) given \( \mathcal{G}_k \) is denoted by
\[
W_{\mathcal{Q}}(\mathcal{X} \mid \mathcal{G}_k) = \sup_{q(\mathcal{X} \mid \mathcal{G}_k) \in \mathcal{Q}} E \{ \log q(\mathcal{X} \mid \mathcal{G}_k) \}.
\]
Any \( q^*(\mathcal{X} \mid \mathcal{G}_k) \) that attains the maximum is conditionally log-optimum given \( \mathcal{G}_k \) and therefore attains the maximum conditional expected growth exponent
\[
w_k^* = E \{ \log q^*(\mathcal{X} \mid \mathcal{G}_k) \mid \mathcal{G}_k \} = \text{ess sup}_{q(\mathcal{X} \mid \mathcal{G}_k) \in \mathcal{Q}} E \{ \log q(\mathcal{X} \mid \mathcal{G}_k) \mid \mathcal{G}_k \}.
\]
We claim that the maximum growth exponent of selections given \( \mathcal{G}_k \) increases to the maximum growth exponent given the limiting \( \sigma \)-field \( \mathcal{G}_\infty \). Thus no gap is left between \( W_{\mathcal{Q}}(\mathcal{X} \mid \mathcal{G}_k) \)
and \( W_\mathcal{Q}(X|\mathcal{G}_\infty) \) in the limit as \( k \to \infty \). This is a general formulation of the no-gap property that was combined with a sandwich argument by Algoet and Cover [2] [4] [3] to prove the Shannon-McMillan-Breiman theorem and a more general ergodic theorem for the maximum growth rate of compounded wealth invested in a stationary market.

**Theorem 11.** If \( \mathcal{G}_k \nearrow \mathcal{G}_\infty \) and if \( -\infty < W_\mathcal{Q}(X|\mathcal{G}_K) \) for some finite \( K \), then

\[
W_\mathcal{Q}(X|\mathcal{G}_k) \nearrow W_\mathcal{Q}(X|\mathcal{G}_\infty) \quad \text{as} \quad k \to \infty. \tag{17}
\]

Suppose that for \( K \leq k \leq \infty \), the maximum growth exponent \( W_\mathcal{Q}(X|\mathcal{G}_k) \) is well defined and finite and attained by a conditionally log-optimum selection \( q^*(x|\mathcal{G}_k) \). Then

\[
\log q^*(X|\mathcal{G}_k) \to \log q^*(X|\mathcal{G}_\infty) \quad \text{in} \quad L^1(P). \tag{18}
\]

Also, \( w_k^* = E\{\log q^*(X|\mathcal{G}_k)|\mathcal{G}_k\} \) is a uniformly integrable submartingale with respect to the \( \sigma \)-fields \( \mathcal{G}_k \), hence \( w_k^* \to w_\infty^* \) almost surely and in \( L^1(P) \).

**Proof:** The family of measurable selections in \( \mathcal{Q} \) given \( \mathcal{G}_k \) is \( \mathcal{G}_k \)-convex and monotonically increasing in \( k \). Thus \( W_\mathcal{Q}(X|\mathcal{G}_k) \) is monotonically increasing to a limit less than or equal to \( W_\mathcal{Q}(X|\mathcal{G}_\infty) \). We show that the limit is equal to \( W_\mathcal{Q}(X|\mathcal{G}_\infty) \).

Pick some \( \gamma(X|\mathcal{G}_K) \) in \( \mathcal{Q} \) given \( \mathcal{G}_K \) such that \( E\{\log \gamma(X|\mathcal{G}_K)\} > -\infty \), and observe that

\[
W_\mathcal{Q}(X|\mathcal{G}_\infty) = E\{\log \gamma(X|\mathcal{G}_K)\} + W_\mathcal{Q}/\gamma(X|\mathcal{G}_\infty),
\]

where

\[
W_\mathcal{Q}/\gamma(X|\mathcal{G}_\infty) = \sup_{q(x|\mathcal{G}_\infty) \in \mathcal{Q}} E\left\{ \log \left( \frac{q(X|\mathcal{G}_\infty)}{\gamma(X|\mathcal{G}_K)} \right) \right\}.
\]

We assume without loss of generality that the convex family \( \mathcal{Q} \) is down-monotone, so that the family of ratios \( q(x|\mathcal{G}_\infty)/\gamma(x|\mathcal{G}_K) \) is a convex down-monotone family of nonnegative measurable functions on \((X, \mathcal{B}) \times (\Omega, \mathcal{G}_\infty)\) containing the constant 1. Now observe that \( \mathcal{G} = \mathcal{G}_k \) is a generating subfield of the \( \sigma \)-field \( \mathcal{G}_\infty = \bigvee_k \mathcal{G}_k \). By the approximation arguments in the proof of Theorem 8, the maximum growth exponent \( W_\mathcal{Q}/\gamma(X|\mathcal{G}_\infty) \) may be taken over measurable selections \( q(x|\mathcal{G}) \) such that \( V(x|\mathcal{G}) = \log[q(x|\mathcal{G})/\gamma(x|\mathcal{G}_K)] \) is a simple function with constant values on the atoms of some finite partition of \( X \times \Omega \) into rectangles \( B \times G \) where \( B \in \mathcal{B} \) and \( G \in \mathcal{G} \). Obviously, \( q(x|\mathcal{G}) = \gamma(x|\mathcal{G}_K) e^{V(x|\mathcal{G})} \) is \( \mathcal{G}_k \)-measurable for some finite \( k \). The desired conclusion (17) follows immediately.

Assertion (18) follows from Theorem 7, because \( E\{\log q^*(X|\mathcal{G}_k)\} = W_\mathcal{Q}(X|\mathcal{G}_k) \) converges to the maximum growth exponent \( E\{\log q^*(X|\mathcal{G}_\infty)\} = W_\mathcal{Q}(X|\mathcal{G}_\infty) \). Notice that \( q^*(x|\mathcal{G}_\infty) \) is conditionally log-optimum given \( \mathcal{G}_\infty \) by Theorem 10. The statements about \( w_k^* \) are justified by the remarks at the end of the previous section.

**F. A water-pouring solution**

It is only in special cases that an explicit formula can be given for the log-optimum selection in a convex family. Here is such a special case which I learned from T. Cover.

Let \( \mathcal{Q} \) be the convex hull of a function \( \gamma(x) \) and all densities \( dQ/dm(x) \) of probability or sub-probability measures \( Q(dx) \) relative to a \( \sigma \)-finite reference measure \( m(dx) \). Thus \( \mathcal{Q} \) is the set of convex combinations \( \beta \gamma(x) + b dQ/dm(x) \) with weights \( \beta, b \geq 0 \) such that \( \beta + b = 1 \). Let \( M(dx) = \gamma(x) m(dx) \), and observe that a convex combination \( \beta^* \gamma(x) + b^* dQ^*/dm(x) \)
is log-optimum in $Q$ iff $\beta^* + b^* \frac{dQ^*}{dM(x)}$ is log-optimum in $Q/\gamma$. The KKT conditions assert that for any sub-probability measure $Q(dx)$,

$$E \left\{ \frac{b^* (dQ - dQ^*)}{\beta^* dM + b^* dQ^*} \right\} \leq 0,$$

and for any weight $b$ in the interval $0 \leq b \leq 1$,

$$E \left\{ \frac{(b - b^*) (dQ^* - dM)}{\beta^* dM + b^* dQ^*} \right\} \leq 0.$$

If $M$ is a probability or sub-probability measure then the log-optimum strategy places the full weight $b^* = 1$ on the log-optimum $Q^*(dx) = P(dx)$ in $Q(m)$. At the other extreme, $\gamma(x)$ is log-optimum in $Q$ and $\beta^* = 1$ iff for every probability measure $Q(dx)$,

$$E_P \left\{ \frac{dQ}{dM} (X) \right\} \leq 1.$$

For this it suffices that $\mathcal{X}$ is a finite or countable set and $M(x) \geq 1$ for all $x \in \mathcal{X}$. Otherwise, both $\beta^*$ and $b^*$ are positive and the log-optimum selection is obtained by water pouring:

$$\frac{dQ^*}{dM}(x) = \frac{1}{b^*} \left[ \frac{dP}{dM}(x) - \beta^* \right]^+,$$

where $\beta^*$ and $b^* = 1 - \beta^*$ satisfy the balance equation

$$E_P \left\{ \frac{dM}{\beta^* dM + b^* dQ^*} \right\} = E_P \left\{ \frac{dQ^*}{\beta^* dM + b^* dQ^*} \right\} = 1.$$

Clearly, $Q^*(dx)$ is supported by the set $\{x : dP/dM(x) \geq \beta^*\}$. The KKT conditions assert that for any subprobability measure $Q(dx)$, the variation $\delta Q = Q - Q^*$ satisfies

$$\int \frac{dP/dM}{\beta^* + [dP/dM - \beta^*]^+} \delta Q \leq 0.$$

This KKT condition holds with equality if $Q(dx)$ is a probability measure supported by $\{x : dP/dM(x) \geq \beta^*\}$, and it holds as a strict inequality otherwise.
IV. LOG-OPTIMUM MEASURES AND TRANSITION KERNELS

For every result about log-optimum selections in a convex family of random variables there is a parallel result about log-optimum selections in a convex family of measures. The counterparts of Theorems 6, 7, 8, 10, 11 will be labeled Theorems 6', 7', 8', 10', 11'.

A. Log-optimum measures

Let X be a random variable with distribution $P(dx)$ on a measurable space $\mathcal{X}$ and let $\mathcal{M}$ be a family of $\sigma$-finite measures on $\mathcal{X}$. Pick a $\sigma$-finite reference measure $\lambda(dx)$ such that $P$ is dominated by $\lambda$, and for any $m \in \mathcal{M}$ let $dm/d\lambda(x)$ denote the density of the absolutely continuous part of $m$ relative to $\lambda$. We say that a measure $m^* \in \mathcal{M}$ is log-optimum in $\mathcal{M}$ (under $P$) if $dm^*/d\lambda(x)$ is log-optimum among all densities $dm/d\lambda(x)$, $m \in \mathcal{M}$. Thus we require that $dm^*/d\lambda(X) > 0$ almost surely and for all alternative $m \in \mathcal{M}$,

$$
E_P \left\{ \log \left( \frac{dm}{dm^*}(X) \right) \right\} = E_P \left\{ \log \left( \frac{d\lambda(X)}{d\lambda^*(X)} \right) \right\} \leq 0.
$$

The concept of log-optimum probability measures arises in maximum likelihood estimation, when the true but unknown distribution $P$ must be inferred from empirical observations. Let $\hat{P}_n$ denote the empirical distribution of independent realizations $X^n = (X_0, X_1, \ldots, X_{n-1})$ of $P$:

$$
\hat{P}_n = \frac{1}{n} \sum_{0 \leq t < n} \delta_{X_t}.
$$

The maximum likelihood estimate of $P$ in a class $\mathcal{M}$ of distributions $M$ based on these $n$ observations is the distribution $\hat{M}_n$ in $\mathcal{M}$ that maximizes the compounded likelihood $dM/d\lambda(X^n) = \prod_{0 \leq t < n} dm/d\lambda(X_t)$ or that maximizes the empirical log-likelihood

$$
E_{\hat{P}_n} \left\{ \log \left( \frac{dM}{d\lambda}(X) \right) \right\} = \frac{1}{n} \sum_{0 \leq t < n} \log \left( \frac{dM}{d\lambda}(X_t) \right).
$$

In other words, the maximum likelihood estimate $\hat{M}_n(dx)$ is the distribution that is log-optimum in the model class $\mathcal{M}$ under the empirical distribution $\hat{P}_n(dx)$.

The conditions for log-optimality of a measure $m^*(dx)$ in a family $\mathcal{M}$ do not depend on the choice of the dominating measure $\lambda(dx)$. In particular, one may choose the true distribution $P$ as reference measure $\lambda$. Thus $m^*$ is log-optimum in $\mathcal{M}$ under $P$ iff $dm^*/dP(X) > 0$ almost surely and for all $m \in \mathcal{M}$,

$$
E_P \left\{ \log \left( \frac{dm}{dP}(X) \right) \right\} \leq 0.
$$

The condition that $dm^*/dP(X) > 0$ almost surely is equivalent to the condition that $P$ is dominated by $m^*$ (notation: $P \ll m^*$). Any log-optimum $m^* \in \mathcal{M}$ attains (if it is well defined) the minimum information

$$
I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} I_m(P) = I_\mathcal{M}(X).
$$

Conversely, if $I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} E\{\log[dP/dm(X)]\}$ is well defined and finite then any $m^* \in \mathcal{M}$ which attains the minimum is log-optimum in $\mathcal{M}$. 

29
If \( P \ll m^* \) and \( E_P\{dm/dm^*(X)\} \leq 1 \) for all \( m \in \mathcal{M} \) then \( m^* \) is log-optimum in \( \mathcal{M} \) by Jensen's inequality (12). The converse is true if \( \mathcal{M} \) is a convex family.

**Theorem 6'.** Let \( X \) be a random variable with distribution \( P \) on a measurable space \( \mathcal{X} \) and let \( \mathcal{M} \) be a convex family of measures on \( \mathcal{X} \). Then \( m^* \in \mathcal{M} \) is log-optimum in \( \mathcal{M} \) iff \( P \ll m^* \) and \( m^* \) satisfies the KKT conditions

\[
E_P\left\{ \frac{dm}{dm^*(X)} \right\} = E_P\left\{ \frac{dm}{dP}(X) \right\} \leq 1, \quad \text{for all } m \in \mathcal{M}.
\]

The existence of a log-optimum \( m^* \) follows from Theorem 7. The measure \( m^* \) need not be unique, but the random variable \( dP/dm^*(X) \) is unique up to almost sure equivalence.

**Theorem 7'.** Let \( \mathcal{M} \) be a convex family of measures on \( \mathcal{X} \) and let \( \bar{\mathcal{M}} \) denote the family of measures \( \bar{m} \) such that \( \log[dm_k/d\bar{m}(X)] \to 0 \) in \( L^1(P) \) for some sequence \( \{m_k\} \) in \( \mathcal{M} \).

(a). The family \( \mathcal{M} \) is convex, \( \mathcal{M} \subseteq \bar{\mathcal{M}} \), and \( I_\mathcal{M}(P) = I_\bar{\mathcal{M}}(P) \).

(b). If \( I_\mathcal{M}(P) \) is well defined and finite then a log-optimum \( m^* \) exists in \( \bar{\mathcal{M}} \). For any \( m_k \in \mathcal{M} \) such that \( I_{m_k}(P) \) converges to the minimum information \( I_\mathcal{M}(P) \), the subprobability \( dQ_k = (dm_k/dm^*)dP \) converges in information to the true distribution \( dP \) and

\[
\log \left( \frac{dP}{dm_k}(X) \right) \to \log \left( \frac{dP}{dm^*}(X) \right) \quad \text{in } L^1(P).
\]

The following result can be used to construct a universal scheme for discriminating distributions \( P \) against measures on a convex family \( \mathcal{M} \). In particular, it can be used to construct a universal test against a composite convex null hypothesis.

**Theorem 8'.** Let \( \mathcal{M} \) be a convex family of \( \sigma \)-finite measures on a measurable space \( \mathcal{X} \) with a countably generated \( \sigma \)-field, and let \( \mu(dx) \) be a particular measure in \( \mathcal{M} \). There exists a sequence \( \{m_k\}_{k \geq 1} \) in \( \mathcal{M} \) such that for every probability distribution \( P \) with \( I_\mu(P) < \infty \),

\[
I_\mathcal{M}(P) = \inf_k I_{m_k}(P).
\]

Proof: For each \( m \in \mathcal{M} \) let \( dm/d\mu(x) \) denote the density of the absolutely continuous part of \( m \) relative to \( \mu \). If \( I_m(P) < \infty \) then \( I_m(P) = I_\mu(P) - E_P\{\log[dm/d\mu(X)]\} \), hence

\[
I_\mathcal{M}(P) = I_\mu(P) - \sup_{m \in \mathcal{M}} E_P\left\{ \log \left( \frac{dm}{d\mu}(X) \right) \right\}.
\]

The family of densities \( \{dm/d\mu(x)\}_{m \in \mathcal{M}} \) is convex and contains the constant \( d\mu/d\mu(x) \equiv 1 \). By Theorem 8, there exists a sequence \( \{m_k\}_{k \geq 1} \) in \( \mathcal{M} \) such that for every distribution \( P \),

\[
\sup_{m \in \mathcal{M}} E_P\left\{ \log \left( \frac{dm}{d\mu}(X) \right) \right\} = \sup_k E_P\left\{ \log \left( \frac{dm_k}{d\mu}(X) \right) \right\}.
\]

If \( I_\mu(P) < \infty \) then

\[
I_\mathcal{M}(P) = I_\mu(P) - \sup_k E_P\left\{ \log \left( \frac{dm_k}{d\mu}(X) \right) \right\} = \inf_k I_{m_k}(P).
\]

This concludes the proof of the theorem. ■
B. Log-optimality in hypothesis testing

Let \( P \) and \( M \) be two probability distributions on a measurable space \( \mathcal{X} \), and suppose we like to test the null hypothesis that a random variable \( X \) has distribution \( P \) against the alternative that \( X \) has distribution \( M \) on the basis of \( n \) independent samples \( X^n = (X_0, \ldots, X_{n-1}) \) of \( X \). A test is specified by an acceptance region \( A_n \subseteq \mathcal{X}^n \). If \( X^n \in A_n \) then we accept the null hypothesis \( P \), otherwise we accept the alternative \( M \). The error probability when \( P \) is true is equal to \( P\{X^n \notin A_n\} \) and the error probability when \( M \) is true is equal to \( M\{X^n \in A_n\} \). According to Stein's lemma, the maximum exponential decay rate of the error probability of the second kind when the error probability of the first kind is bounded away from 1 is equal to the divergence \( I_M(P) \), and the maximum decay rate is attained by the likelihood ratio test of \( P \) against \( M \). For a precise statement and proof of Stein's lemma see Section 12.8 of Cover and Thomas [27].

Now we test \( P \) against a composite alternative \( \mathcal{M} \) (it is assumed that \( P \notin \mathcal{M} \)). The log-optimum \( M^* \) is the worst alternative when discriminating \( P \) against probability measures in the family \( \mathcal{M} \). Indeed, the error probability of the second kind when testing \( P \) against a simple alternative \( M \) in \( \mathcal{M} \) decays exponentially with maximum rate \( I_M(P) \), and this maximum decay rate is minimum when we test \( P \) against the log-optimum \( M^* \) in \( \mathcal{M} \) because \( M^* \) attains \( I_M(P) = \inf_{M \in \mathcal{M}} I_M(P) \). We now assume that \( \mathcal{M} \) is convex and prove that the likelihood ratio test of \( P \) against \( M^* \) has error exponent at least \( I_M(P) \) when we test \( P \) against any alternative in \( \mathcal{M} \). Thus one single test attains at least the minimax error exponent \( I_M(P) \) against every alternative in the family \( \mathcal{M} \).

**Theorem 12.** Let \( \mathcal{M} \) be a convex family of probability distributions on \( \mathcal{X} \), and suppose \( M^*(dx) \) is log-optimum in \( \mathcal{M} \) under the distribution \( P \) on \( \mathcal{X} \). Fix \( 0 < \epsilon < 1 \) and consider the likelihood ratio test with acceptance region

\[
A_n^\epsilon = \left\{ \frac{1}{n} \log \left( \frac{dP}{dM^*}(X^n) \right) \geq d_n^\epsilon \right\} = \left\{ \frac{dP}{dM^*}(X^n) \geq e^{nd_n^\epsilon} \right\},
\]

where \( d_n^\epsilon \) is the supremum of all values \( d \) such that \( P\{dP/dM^*(X^n) \leq e^{nd}\} \leq \epsilon \). Then the error probability \( M^*\{X^n \in A_n^\epsilon\} \) decays exponentially with rate \( I_M(P) \):

\[
-\frac{1}{n} \log M^*\{X^n \in A_n^\epsilon\} \to I_M(P) = I_M(P).
\]  

(19)

For any alternative \( M \in \mathcal{M} \), the error probability \( M\{X^n \in A_n^\epsilon\} \) decays exponentially with limiting rate at least \( I_M(P) \). In fact,

\[
-\frac{1}{n} \log \left( \sup_{M \in \mathcal{M}} M\{X^n \in A_n^\epsilon\} \right) \to I_M(P).
\]  

(20)

**Proof:** Assertion (19) follows from Stein's lemma. If \( M \in \mathcal{M} \) then by the Markov inequality and the Kuhn-Tucker conditions for log-optimality of \( M^*(dx) \),

\[
ed^{nd_n^\epsilon} M\{X^n \in A_n^\epsilon\} \leq \int_{A_n^\epsilon} \frac{dP}{dM^*}(x^n) M(dx^n) \leq \int \frac{dM}{dM^*}(x^n) P(dx^n) \leq 1
\]

and consequently

\[
\sup_{M \in \mathcal{M}} M\{X^n \in A_n^\epsilon\} \leq e^{-nd_n^\epsilon}.
\]

31
Assertion (20) follows since \( d_n^* \rightarrow I_M(P) \) by the Shannon-McMillan theorem.

The conclusions (19) and (20) of the theorem remain valid when the measures in \( M \) are not normalized. However, there is then no hypothesis testing interpretation.

**C. Conditionally log-optimum transition kernels**

Suppose the random variable \( X \) is defined on a probability space \((\Omega, \mathcal{F}, P)\) and \( \mathcal{G} \) is a sub-\( \sigma \)-field of \( \mathcal{F} \). For any family \( \Pi \) of transition kernels \( \mu(dx|\mathcal{G}) \) from \((\Omega, \mathcal{G})\) to \((X, \mathcal{B})\) we define the minimum conditional information

\[
I_{\Pi}(X|\mathcal{G}) = \inf_{\mu(dx|\mathcal{G}) \in \Pi} I_{\mu}(X|\mathcal{G}).
\]

Let \( \bar{\Pi} \) denote the family of transition kernels \( \bar{\mu}(dx|\mathcal{G}) \) such that for some \( \mu_k(dx|\mathcal{G}) \) in \( \Pi \),

\[
\log \left( \frac{d\mu_k}{d\bar{\mu}}(X|\mathcal{G}) \right) = \log \left( \frac{d\mu_k}{dP}(X|\mathcal{G}) / \frac{d\bar{\mu}}{dP}(X|\mathcal{G}) \right) \rightarrow 0 \quad \text{in} \quad L^1(P).
\]

If \( \Pi \) is convex then \( \bar{\Pi} \) is convex, \( \Pi \subseteq \bar{\Pi} \), and \( I_{\Pi}(X|\mathcal{G}) = I_{\bar{\Pi}}(X|\mathcal{G}) \). If \( I_{\Pi}(X|\mathcal{G}) \) is well defined and finite, then there exists a log-optimum transition kernel \( \mu^*(dx|\mathcal{G}) \) in \( \bar{\Pi} \) such that for any \( \mu_k(dx|\mathcal{G}) \) in \( \Pi \) with \( I_{\mu_k}(X|\mathcal{G}) \) converging to the infimum \( I_{\bar{\Pi}}(X|\mathcal{G}) \), we have

\[
\log \left( \frac{d\mu_k}{dP}(X|\mathcal{G}) \right) \rightarrow \log \left( \frac{d\mu^*}{dP}(X|\mathcal{G}) \right) \quad \text{in} \quad L^1(P).
\]

The conditional likelihood \( d\mu^*/dP(X|\mathcal{G}) \) is uniquely characterized up to almost sure equivalence by the KKT conditions which assert that for any \( \mu(dx|\mathcal{G}) \) in \( \Pi \),

\[
E \left\{ \frac{d\mu}{d\mu^*}(X|\mathcal{G}) \right\} = E \left\{ \frac{d\mu}{dP}(X|\mathcal{G}) / \frac{d\mu^*}{dP}(X|\mathcal{G}) \right\} \leq 1.
\]

Let \( \mathcal{H} \) be a sub-\( \sigma \)-field of \( \mathcal{G} \) (e.g. \( \mathcal{H} = \mathcal{G} \)). We say that \( \Pi \) is \( \mathcal{H} \)-convex if \( \Pi \) is closed under taking convex combinations with \( \mathcal{H} \)-measurable weights.

**Theorem 10**. If \( \Pi \) is \( \mathcal{H} \)-convex then \( \bar{\Pi} \) is \( \mathcal{H} \)-convex as well, and the log-optimum \( \mu^*(dx|\mathcal{G}) \) in \( \bar{\Pi} \) satisfies the conditional KKT conditions

\[
E \left\{ \frac{d\mu}{d\mu^*}(X|\mathcal{G}) | \mathcal{H} \right\} = E \left\{ \frac{d\mu}{dP}(X|\mathcal{G}) / \frac{d\mu^*}{dP}(X|\mathcal{G}) | \mathcal{H} \right\} \leq 1 \quad \text{almost surely.}
\]

For each \( k \) let \( \Pi_k \) be a convex family of transition kernels \( \mu(dx|\mathcal{G}) \) from \((\Omega, \mathcal{G})\) to \((X, \mathcal{B})\), and suppose \( \Pi_k \) is monotonically increasing to the limiting family \( \Pi_\infty = \bigcup_k \Pi_k \) as \( k \rightarrow \infty \). If \( \infty > I_{\Pi_k}(X|\mathcal{G}) \) for some finite \( K \), then

\[
I_{\Pi_k}(X|\mathcal{G}) \searrow I_{\Pi_\infty}(X|\mathcal{G}) \quad \text{as} \quad k \rightarrow \infty.
\]

Suppose that for all \( K \leq k \leq \infty \), \( I_{\Pi_k}(X|\mathcal{G}) \) is finite and \( \mu^*_k(dx|\mathcal{G}) \) is log-optimum in the closure \( \bar{\Pi}_k \) of the family \( \Pi_k \). Then

\[
\log \left( \frac{dP}{d\mu^*_k}(X|\mathcal{G}) \right) \rightarrow \log \left( \frac{dP}{d\mu^*_\infty}(X|\mathcal{G}) \right) \quad \text{in} \quad L^1(P).
\]
The following result for transition kernels is the dual of Theorem 11 for random variables.

**Theorem 11'**. Let \( \mathcal{G}_k \) be a sub-\( \sigma \)-field of \( \mathcal{G} \) such that \( \mathcal{G}_k \) is monotonically increasing to the limiting \( \sigma \)-field \( \mathcal{G}_\infty \) as \( k \to \infty \). If \( \Pi_k \) is \( \mathcal{G}_k \)-convex for all finite \( k \), then the log-optimum \( \mu^*(dx|\mathcal{G}_k) \) in \( \Pi_k \) is conditionally log-optimum given \( \mathcal{G}_k \) for all \( k \) including \( k = \infty \). Also, the minimum conditional expected information

\[
i^*_k = \underline{\text{ess inf}}_{\mu_k(dx|\mathcal{G}) \in \Pi_k} E \left\{ \log \left( \frac{dP}{d\mu_k}(X|\mathcal{G}) \right) \middle| \mathcal{G}_k \right\} = E \left\{ \log \left( \frac{dP}{d\mu^*_k}(X|\mathcal{G}) \right) \middle| \mathcal{G}_k \right\}
\]

is a uniformly integrable supermartingale with respect to the \( \sigma \)-fields \( \mathcal{G}_k \) and

\[
i^*_k \to i^*_\infty \quad \text{in } L^1(P).
\]

Let \( \mathcal{M} \) be a convex family of measures on \( \mathcal{X} \) and for each \( k \) let \( \Pi_k \) denote the \( \mathcal{G}_k \)-convex family of measurable selections \( \mu(dx|\mathcal{G}_k) \) in \( \mathcal{M} \) given \( \mathcal{G}_k \). The theorem is applicable when \( \mathcal{G} = \mathcal{G}_\infty \) and each \( \Pi_k \) is regarded as a \( \mathcal{G}_k \)-convex family of transition kernels from \( (\Omega, \mathcal{G}) \) to \( (\mathcal{X}, \mathcal{B}) \). On the other hand, if \( I_{\Pi_k}(X|\mathcal{G}_K) \) is finite for some finite \( K \) then

\[
I_{\Pi_k}(X|\mathcal{G}_k) \nearrow I_{\Pi_k}(X|\mathcal{G}_\infty) \quad \text{as } k \to \infty.
\]

Suppose the conditional mutual information \( I(X; \mathcal{G}_\infty|\mathcal{G}_K) \) is finite for some finite \( K \). Then \( I(X; \mathcal{G}_\infty|\mathcal{G}_k) \to 0 \) as \( k \to \infty \), and using (21) we may conclude that

\[
I_{\Pi_k}(X|\mathcal{G}_k) = I_{\Pi_k}(X|\mathcal{G}_\infty) - I(X; \mathcal{G}_\infty|\mathcal{G}_k) \to I_{\Pi_\infty}(X|\mathcal{G}_\infty) \quad \text{as } k \to \infty.
\]

The convergence need not be monotone, but it can be shown that convergence of the expectations to a finite limit implies mean convergence

\[
\log \left( \frac{dP}{d\mu_k}(X|\mathcal{G}_k) \right) \to \log \left( \frac{dP}{d\mu^*_\infty}(X|\mathcal{G}_\infty) \right) \quad \text{in } L^1(P).
\]

**D. Explicit representation of log-optimum measures**

Generally speaking, the log-optimum \( M^*(dx) \) in a convex family \( \mathcal{M} \) of probability measures on \( \mathcal{X} \) is only implicitly characterized by the KKT conditions. We discuss some special cases where an explicit representation of \( M^*(dx) \) is possible. In those special cases, \( M^* \) is optimal not only according to the logarithmic criterion but according to a variety of other criteria as well.

Let \( X \) be a random variable with distribution \( P(dx) \) on a measurable space \( \mathcal{X} \) and suppose \( P(dx) \) is dominated by a \( \sigma \)-finite reference measure \( \lambda(dx) \) on \( \mathcal{X} \). Let \( \mathcal{B} \) be a \( \sigma \)-field of measurable subsets of \( \mathcal{X} \), and let \( \mathcal{M} \) denote the family of all \( \lambda \)-dominated probability measures \( M(dx) \) such that the density \( dM/d\lambda(x) \) is \( \mathcal{B} \)-measurable. Then \( \mathcal{M} \) is both a convex (in fact, \( \mathcal{B} \)-convex) family and an exponential family. This exponential family is orthogonal to the convex family \( \Pi \) of all probability measures on \( \mathcal{X} \) that have the same restriction to the sub-\( \sigma \)-field \( \mathcal{B} \) as the distribution \( P \). The log-optimum \( M^*(dx) \) in \( \mathcal{M} \) is the unique distribution in \( \Pi \cap \mathcal{M} \) and has density equal to the conditional expectation

\[
\frac{dM^*}{d\lambda} = E^\lambda \left\{ \frac{dP}{d\lambda} \middle| \mathcal{B} \right\}.
\]
(The operator $E_\lambda \{ \cdot | B \}$ is the continuous extension to $L^1(\lambda)$ of the operator that maps any square integrable function in $L^2(\lambda)$ to its orthogonal projection on the linear subspace of $B$-measurable functions.) In fact, $M^*(dx)$ is the $I$-projection of every distribution $M(dx)$ in $\mathcal{M}$ on the convex family $\Pi$. For all $M(dx)$ in $\mathcal{M}$, we have the Pythagorean decomposition

$$I(P|M) = I(P|M^*) + I(M^*|M).$$

Similar results hold when the $\sigma$-field $B$ is replaced by a $\sigma$-lattice $\mathcal{U}$ of measurable subsets of $\mathcal{X}$. Thus we assume that $\mathcal{U}$ contains the empty set $\emptyset$ and the set $\mathcal{X}$, and $\mathcal{U}$ is closed under countable unions and countable intersections but not necessarily under complementation. Now, $\mathcal{M}$ denotes the set of $\lambda$-dominated probability measures $M(dx)$ such that the density $dM/d\lambda(x)$ is $\mathcal{U}$-measurable. A real valued function $f(x)$ is called $\mathcal{U}$-measurable if for every real number $a$, the set $\{ x : f(x) \geq a \}$ belongs to $\mathcal{U}$. It is equivalent to require that $\{ x : f(x) > a \}$ belongs to $\mathcal{U}$ for all real $a$. The cone generated by $\mathcal{M}$ is a convex cone and a lattice, and is contained in the cone $\mathcal{K}$ of all $\lambda$-dominated measures $M(dx)$ whose density $dM/d\lambda(x)$ is $\mathcal{U}$-measurable.

The conditional expectation $E_\lambda \{ dP/d\lambda | \mathcal{U} \}$ given the $\sigma$-lattice $\mathcal{U}$ was defined by Brunk and Johansen [22] as a generalized Radon-Nikodym derivative. If $dP/d\lambda$ is a square integrable function in $L^2(\lambda)$, then $E_\lambda \{ dP/d\lambda | \mathcal{U} \}$ is simply the orthogonal projection of the density $dP/d\lambda$ on the closed convex cone of $\mathcal{U}$-measurable functions in $L^2(\lambda)$.

Let $M^*(dx)$ denote the probability measure whose density $dM^*/d\lambda$ is equal to the conditional expectation $E_\lambda \{ dP/d\lambda | \mathcal{U} \}$. From many points of view, $M^*(dx)$ is the best approximation of $P(dx)$ in the convex set $\mathcal{M}$ and even in the convex cone $\mathcal{K}$. Indeed, for any convex function $\Phi(p)$ with left derivative $\phi(p)$ and for any real numbers $p, m$ let

$$\Delta^\Phi(p|m) = \Phi(p) - \Phi(m) - (p - m)\phi(m).$$

For finite measures $P(dx)$ and $M(dx)$ we consider the divergence

$$D^\Phi(P|M) = \int \Delta^\Phi \left( \frac{dP}{d\lambda} \bigg| \frac{dM}{d\lambda} \right) d\lambda.$$

Notice that $D^\Phi(P|M) \geq 0$ with equality if $P = M$. Brunk [23] has shown that the following Pythagorean inequality holds for every finite measure $M(dx)$ in the convex cone $\mathcal{K}$:

$$D^\Phi(P|M) \geq D^\Phi(P|M^*) + D^\Phi(M^*|M).$$

Thus $M^*(dx)$ attains the minimum divergence $D^\Phi(P|M^*) = \inf_{M \in \mathcal{M}} \ D^\Phi(P|M)$ according to any convex criterion $\Phi$. If $\Phi(p) = p^2/2$ then $\Delta^\Phi(p|m) = (p - m)^2/2$, and if $dP/d\lambda$ is a function in $L^2(\lambda)$ then $E_\lambda \{ dP/d\lambda | \mathcal{U} \}$ is the least squares approximation of $dP/d\lambda$ among all $\mathcal{U}$-measurable random variables in $L^2(\lambda)$. If $\Phi(p) = p \log p$ then $\phi(p) = 1 + \log p$ and

$$\Delta^\Phi(p|m) = p \log \left( \frac{p}{m} \right) - p + m,$$

$$D^\Phi(P|M) = I(P|M) - P(\mathcal{X}) + M(\mathcal{X}).$$

The distribution $M^*(dx)$ with density $dM^*/d\lambda = E_\lambda \{ dP/d\lambda | \mathcal{U} \}$ is log-optimum in $\mathcal{M}$ since for any alternative probability measure $M(dx)$ in $\mathcal{M}$, we have the Pythagorean inequality

$$I(P|M) \geq I(P|M^*) + I(M^*|M).$$
The concept of σ-lattice is closely related to the concept of quasi-order. A quasi-order on the set \( \mathcal{X} \) is a reflexive transitive relation on \( \mathcal{X} \). Such a relation is a partial order if it is antisymmetric, an equivalence relation if it is symmetric. A real-valued function \( f(x) \) is called isotonic with respect to a quasi-order \( \preceq \) on \( \mathcal{X} \) if it is order-preserving \( (x \preceq x' \Rightarrow f(x) \leq f(x')) \). A subset \( U \subseteq \mathcal{X} \) is called an upper set for the quasi-order \( \preceq \) if its indicator function is isotonic (if \( x \preceq x' \) and \( x \in U \), then \( x' \in U \)). The measurable subsets of \( \mathcal{X} \) that are upper sets form a σ-lattice \( \mathcal{U} \), and the measurable functions on \( \mathcal{X} \) that are isotonic are exactly those that are \( \mathcal{U} \)-measurable. The conditional expectation \( E_X \{ dP/d\lambda \mid U \} \) is often called the isotonic regression of \( dP/d\lambda \) because it is the best approximation of \( dP/d\lambda \) among all isotonic measurable functions on \( \mathcal{X} \).

Much of what is known about isotonic regression and conditional expectations given a σ-lattice is compiled in the textbook on order-restricted inference of Barlow, Bartholomew, Bremner and Brunk [6]. See also Robertson, Wright and Dijkstra [107]. In general, one may consider the problem of finding the best approximation of a given integrable function \( p(x) \) among all measurable functions \( f(x) \) such that \( f(x') - f(x) \) is constrained to lie in some interval \( [l(x, x'), u(x, x')] \) whenever \( x \preceq x' \) is a comparable pair for the quasi-order. The set of measurable functions \( f(x) \) that satisfy the constraints is a σ-lattice of measurable functions on \( \mathcal{X} \), and the best approximation of \( p(x) \) in this σ-lattice is simply the conditional expectation of \( p(x) \) given this σ-lattice. The solution can be described graphically as a taut string if the quasi-order is a linear order. The Fenchel dual problem is a network flow problem, and it has been pointed out by Veinott [124] that the invariance of the optimal solution under the choice of optimization criterion is also valid for the dual network flow problem. See also Barlow and Brunk [7] and Jewell [67].

E. Orthoprojections and I-projections

There is an interesting duality between minimizing \( I(P|M) \) over all \( M \) in a geodesically convex family \( \mathcal{M} \) and minimizing \( I(P|M) \) over all \( P \) in a convex family \( \Pi \). The first problem arises in maximum likelihood estimation, the second in large deviation theory. The parallelism between the two problems is explained in Section 22 of Čencov [29].

In maximum likelihood estimation, we observe independent samples \( X^n = (X_0, \ldots, X_{n-1}) \) of an unknown distribution \( P \) and we find the distribution \( \hat{M}_n \) in some model class \( \mathcal{M} \) that is closest to the empirical distribution \( \hat{P}_n \) in information divergence. If the models in the class \( \mathcal{M} \) are sufficiently regular, then the maximum likelihood estimate \( \hat{M}_n \) will converge to the distribution \( M^\ast \) in \( \mathcal{M} \) which attains the minimum divergence \( I(P|M) = \inf_{M \in \mathcal{M}} I(P|M) \). (We assume that \( M^\ast \) is unique.) For a regular parametric model class \( \mathcal{M} = \{ M_\theta \}_{\theta \in \Theta} \), the maximum likelihood estimate \( \hat{\theta}_n \) will converge to the parameter value \( \theta^\ast \) of the distribution \( M_{\theta^\ast} = M^\ast \) that attains the minimum \( I_M(P) \). This was proved by Wald [126] in case \( P \) is a member of \( \mathcal{M} \) (hence \( P = M^\ast \)), and by White [127] in general.

Suppose the distributions in \( \mathcal{M} \) are mutually absolutely continuous. The family \( \mathcal{M} \) is called geodesically convex if for arbitrary \( M_0(dx) \) and \( M_1(dx) \) in \( \mathcal{M} \) and for \( 0 < \epsilon < 1 \), \( \mathcal{M} \) contains the weighted geodesic mean

\[
M_\epsilon(dx) = \frac{M_0(dx)^{1-\epsilon}M_1(dx)^\epsilon}{\int M_0(dx)^{1-\epsilon}M_1(dx)^\epsilon}.
\]

Suppose \( \mathcal{M} \) is geodesically convex and \( I(P|M) \) is finite. Čencov [28] [29] proved that if some distribution \( M^\ast \in \mathcal{M} \) attains the minimum divergence \( I(P|M) = \inf_{M \in \mathcal{M}} I(P|M) \),
then $M^*$ is unique and the following Pythagorean inequality holds for all $M \in \mathcal{M}$:

$$I(P|M) \geq I(P|M^*) + I(M^*|M).$$

(22)

Čencov called $M^*$ the orthoprojection of $P$ on $\mathcal{M}$. Any $M_k \in \mathcal{M}$ such that $I(P|M_k)$ converges to the minimum $I(P|\mathcal{M})$ will converge to $M^*$ in information in the sense that $I(M^*|M_k) \to 0$ (because $0 \leq I(M^*|M_k) \leq I(P|M_k) - I(P|M^*)$ converges to zero).

If $\mathcal{M}$ is convex, then every sequence $M_k \in \mathcal{M}$ such that $I(P|M_k)$ converges to the minimum divergence $I(P|\mathcal{M})$ converges to $M^*$ in information in the sense that $I(M^*|M_k) \to 0$. For an arbitrary convex family $\mathcal{M}$, there is no guarantee that the Pythagorean inequality (22) will hold for all $M \in \mathcal{M}$. However, it will if $\mathcal{M}$ is both convex and geodesically convex. An example of a family $\mathcal{M}$ which is both convex and geodesically convex is the family of distributions that are equivalent to $P$ and whose density is isotonic with respect to some quasi-order on $\mathcal{X}$ or measurable with respect to some $\sigma$-lattice of measurable subsets of $\mathcal{X}$. Another example is the set of distributions that are equivalent to $P$ and invariant under some Markov transition kernel from $\mathcal{X}$ to $\mathcal{X}$ (see Section 18.7 of Čencov [39] for the proof).

Sanov’s theorem in large deviation theory is about the asymptotic behavior of the empirical distribution $\hat{P}_n$ of $n$ independent samples $X^n = (X_1, \ldots, X_{n-1})$ of a probability distribution $M$ on the space $\mathcal{X}$. If $\Pi$ is a closed convex set of probability measures on $\mathcal{X}$ with nonempty interior, then the probability $M\{\hat{P}_n \in \Pi\}$ decreases exponentially fast with limiting rate equal to the minimum divergence $I(\Pi|\mathcal{M}) = \inf_{P \in \Pi} I(P|\mathcal{M})$. Thus $I_M(P) = I(P|\mathcal{M})$ is the exponential decay rate of the probability that $\hat{P}_n$ will deviate significantly from the true distribution $M$ and fall in a small convex neighborhood of the alternative distribution $P$. By Stein’s lemma, $I(P|\mathcal{M})$ is the maximum exponential decay rate of the error probability when we try to discriminate $P$ against the null hypothesis $M$ on the basis of samples of $M$ using tests that correctly accept the alternative $P$ given sufficiently many samples of $P$.

Suppose $\Pi$ is convex and $I(\Pi|\mathcal{M})$ is finite. The $I$-projection of $M$ on the family $\Pi$ is defined as the distribution $P^*$ in $\Pi$ that attains the minimum divergence $I(\Pi|\mathcal{M})$. Csiszár [35] proved that this $I$-projection exists if $\Pi$ is closed in variation norm, and that for any distribution $P \in \Pi$ we have the Pythagorean inequality

$$I(P|\mathcal{M}) \geq I(P|P^*) + I(P^*|\mathcal{M}).$$

Equality holds if $P^*$ falls in the algebraic interior of $\Pi$. Any sequence $P_k \in \Pi$ for which $I(P_k|\mathcal{M})$ converges to the minimum $I(\Pi|\mathcal{M})$ converges to $P^*$ in information in the sense that $I(P_k|P^*) \to 0$ (because $0 \leq I(P_k|P^*) \leq I(P_k|\mathcal{M}) - I(P^*|\mathcal{M})$ converges to zero). Csiszár’s result for the $I$-projection of $M$ on a convex family $\Pi$ is analogous to Čencov’s result for the orthoprojection of $P$ on a geodesically convex family $\mathcal{M}$.
V. POLAR SETS, CONVEX CORNERS, AND THE MINIMAX PROPERTY

A. Polar sets and convex corners

Let $X$ be a random variable with distribution $P$ on the space $\mathcal{X}$. If $q(x) \geq 0$ is a function and $m(dx)$ is a measure on $\mathcal{X}$ such that $\int q dm \leq 1$, then the growth exponent $W_q(P) = E_P\{\log q(X)\}$ cannot exceed the information $I_m(P) = E_P\{\log[dP/dm(X)]\}$ by Theorem 1. This inequality continues to hold when $W_q(P)$ is maximized over a family $Q$ and $I_m(P)$ is minimized over a family $\mathcal{M}$ such that $\int q dm \leq 1$ for all $q \in Q$ and $m \in \mathcal{M}$.

**Theorem 13.** Let $Q$ be a family of nonnegative measurable functions and let $\mathcal{M}$ be a family of measures on $\mathcal{X}$ such that the polar condition $\int q dm \leq 1$ holds for all $q \in Q$ and $m \in \mathcal{M}$. If $W_q(P) > -\infty$ for some $q \in Q$ or $I_m(P) < \infty$ for some $m \in \mathcal{M}$, then

$$W_Q(P) = \sup_{q \in Q} W_q(P) \leq \inf_{m \in \mathcal{M}} I_m(P) = I_{\mathcal{M}}(P).$$

The gap between $W_Q(P)$ and $I_{\mathcal{M}}(P)$ is equal to the minimum divergence of the true distribution $dP$ relative to subprobability measures $dQ = q dm$ where $q \in Q$ and $m \in \mathcal{M}$:

$$\inf_{dQ=q dm, \quad q \in Q, \quad m \in \mathcal{M}} I_Q(P) = \inf_{m \in \mathcal{M}} I_m(P) - \sup_{q \in Q} W_q(P) = I_{\mathcal{M}}(P) - W_Q(P).$$

The infimum is attained by $q^* dm^*$ iff $q^*$ is log-optimum in $Q$ and $m^*$ is log-optimum in $\mathcal{M}$, and the inequality $I_{\mathcal{M}}(P) - W_Q(P) \geq 0$ holds with equality iff $q^* dm^* = dP$.

The gap between the maximum growth exponent $W_Q(P)$ and the minimum information functional $I_{\mathcal{M}}(P)$ can only decrease if the family $Q$ and/or the family $\mathcal{M}$ is enlarged. We shall prove that the gap vanishes when $Q$ and $\mathcal{M}$ are maximal subject to the constraint that $\int q dm \leq 1$ for all $q \in Q$ and $m \in \mathcal{M}$.

For any family $Q$ of measurable functions $q(x) \geq 0$ we define the polar family of measures

$$\mathcal{M}(Q) = \{m(dx) : \int q dm \leq 1 \text{ for all } q \in Q\},$$

and for any family $\mathcal{M}$ of measures $m(dx)$ we define the polar family of measurable functions

$$Q(\mathcal{M}) = \{q(x) : q(x) \geq 0, \int q dm \leq 1 \text{ for all } m \in \mathcal{M}\}.$$

The convex corners generated by $Q$ and $\mathcal{M}$ are defined as the polars of the polars and are denoted by $\hat{Q}$ and $\hat{\mathcal{M}}$, respectively:

$$\hat{Q} = Q(\mathcal{M}(Q)), \quad \hat{\mathcal{M}} = \mathcal{M}(Q(\mathcal{M})).$$

The mappings $Q \mapsto \mathcal{M}(Q)$ and $\mathcal{M} \mapsto Q(\mathcal{M})$ define order reversing correspondences between families of functions and families of measures on $\mathcal{X}$:

$$Q_1 \subseteq Q_2 \Rightarrow \mathcal{M}(Q_1) \supseteq \mathcal{M}(Q_2),$$

$$\mathcal{M}_1 \subseteq \mathcal{M}_2 \Rightarrow Q(\mathcal{M}_1) \supseteq Q(\mathcal{M}_2).$$

The mapping $Q \mapsto \hat{Q}$ satisfies the axioms for a closure operator:

$$Q \subseteq \hat{Q}, \quad \hat{Q} = \hat{\hat{Q}}, \quad Q_1 \subseteq Q_2 \Rightarrow \hat{Q}_1 \supseteq \hat{Q}_2.$$
Similarly $\mathcal{M} \mapsto \hat{\mathcal{M}}$ is a closure operator, and furthermore

$$\mathcal{M}(\mathcal{Q}) = \mathcal{M}(\hat{\mathcal{Q}}), \quad \mathcal{Q}(\mathcal{M}) = \mathcal{Q}(\hat{\mathcal{M}}).$$

The class of convex corners of functions on $\mathcal{X}$ is a complete lattice, and so is the class of convex corners of measures on $\mathcal{X}$. The correspondences $\mathcal{Q} \mapsto \mathcal{M}(\mathcal{Q})$ and $\mathcal{M} \mapsto \mathcal{Q}(\mathcal{M})$ define a Galois connection between these two lattices. See Chapter IV of Birkhoff [16] for more on polarities, closure operators, complete lattices, and Galois connections.

We say that a family $\mathcal{Q}$ of functions on $\mathcal{X}$ is a convex corner if $\mathcal{Q} = \mathcal{Q}(\mathcal{M})$ for some family $\mathcal{M}$ of measures on $\mathcal{X}$, or equivalently if $\mathcal{Q} = \hat{\mathcal{Q}}$. Similarly, a family $\mathcal{M}$ of measures on $\mathcal{X}$ is called a convex corner if $\mathcal{M} = \mathcal{M}(\mathcal{Q})$ for some family $\mathcal{Q}$ of functions, or equivalently if $\mathcal{M} = \hat{\mathcal{M}}$. In general $\hat{\mathcal{Q}}$ is the smallest convex corner containing $\mathcal{Q}$, and $\hat{\mathcal{M}}$ is the smallest convex corner containing $\mathcal{M}$.

One may define a Galois correspondence between the exposed faces of a convex corner $\mathcal{Q}$ and the exposed faces of the polar convex corner $\mathcal{M}$. For subsets $\Gamma \subseteq \mathcal{Q}$ and $\Pi \subseteq \mathcal{M}$ let

$$\Gamma^d = \{ m \in \mathcal{M} : \int q \, dm = 1 \text{ for all } q \in \mathcal{Q} \},$$

$$\Pi^d = \{ q \in \mathcal{Q} : \int q \, dm = 1 \text{ for all } m \in \mathcal{M} \}.$$  

Note that $\Gamma^d = \Gamma^{d\hat{\mathcal{Q}}}$ and $\Pi^d = \Pi^{d\hat{\mathcal{M}}}$. The set $\Gamma^{d\mathcal{Q}}$ is the smallest exposed face of $\mathcal{Q}$ containing $\Gamma$ and $\Pi^{d\mathcal{M}}$ is the smallest exposed face of $\mathcal{M}$ containing $\Pi$.

We say that a set of nonnegative measurable functions on $\mathcal{X}$ is completely convex if the barycenter of any probability distribution on that set also belongs to that set. We similarly define completely convex families of measures on $\mathcal{X}$.

**Theorem 14.** Any convex corner is down-monotone, completely convex, and closed under taking lower limits. In particular, a convex corner is closed under pointwise limits.

**Proof:** It is obvious that $\mathcal{Q}(\mathcal{M})$ and $\mathcal{M}(\mathcal{Q})$ are down-monotone. If $\nu(dm)$ is a probability distribution on $\mathcal{M}(\mathcal{Q})$ with barycenter $m_{\nu} = \int_{\mathcal{M}(\mathcal{Q})} m \, \nu(dm)$, then $m_{\nu} \in \mathcal{M}(\mathcal{Q})$ since

$$\int q \, dm_{\nu} = \int_{\mathcal{M}(\mathcal{Q})} \int_{\mathcal{X}} q(x) \, m_{\nu} \, dx \, \nu(dm) \leq \int_{\mathcal{M}(\mathcal{Q})} \nu(dm) = 1, \quad \text{for all } q \in \mathcal{Q}.$$  

Similarly, if $\nu(dq)$ is a probability distribution on $\mathcal{Q}(\mathcal{M})$ with barycenter $q_{\nu} = \int_{\mathcal{Q}(\mathcal{M})} q \, \nu(dq)$, then $q_{\nu} \in \mathcal{Q}(\mathcal{M})$ since

$$\int q_{\nu} \, dm = \int_{\mathcal{Q}(\mathcal{M})} \int_{\mathcal{X}} q_{\nu} \, m \, dx \, \nu(dq) \leq \int_{\mathcal{Q}(\mathcal{M})} \nu(dq) = 1, \quad \text{for all } m \in \mathcal{M}.$$  

If $\{q_k\}$ is a sequence in $\mathcal{Q}(\mathcal{M})$ then by Fatou’s lemma,

$$\int \liminf_k q_k \, dm \leq \liminf_k \int q_k \, dm \leq 1, \quad \text{for all } m \in \mathcal{M}.$$  

The lower limit of any sequence in $\mathcal{Q}(\mathcal{M})$ also belongs to $\mathcal{Q}(\mathcal{M})$, so $\mathcal{Q}(\mathcal{M})$ is closed under pointwise convergence. If $\{m_k\}$ is any sequence of measures on $\mathcal{X}$ then we define

$$\liminf_k m_k(dx) = \lambda(dx) \cdot \liminf_k \frac{dm_k}{d\lambda}(x)$$

where $\lambda(dx)$ is any $\sigma$-finite dominating measure on $\mathcal{X}$. The lower limit of any sequence in $\mathcal{M}(\mathcal{Q})$ also belongs to $\mathcal{M}(\mathcal{Q})$, by Fatou’s lemma. 

38
Given a convex family \( \mathcal{Q} \), one may wonder how the convex corner \( \hat{\mathcal{Q}} \) is related to the closure \( \bar{\mathcal{Q}} \) which is defined in Theorem 7. The family \( \hat{\mathcal{Q}} \) depends on the distribution \( P \) whereas \( \bar{\mathcal{Q}} \) is defined without reference to \( P \). If \( \tilde{q}(x) \) is a function in \( \mathcal{Q} \) then there exists a sequence of functions \( q_k(x) \) in \( \mathcal{Q} \) such that \( \log[q_k(X)/\tilde{q}(X)] \to 0 \) in \( L^1(P) \), and by passing to a subsequence we may assume that \( q_k(X) \to q(X) \) \( P \)-almost surely. Thus for every function \( \tilde{q}(x) \) in \( \mathcal{Q} \) there exists a function \( \hat{q}(x) = \lim \inf_k q_k(x) \) in \( \hat{\mathcal{Q}} \) such that \( \hat{q}(X) = \tilde{q}(X) \) \( P \)-almost surely. This means that the family of random variables \( \{q(X) : q \in \mathcal{Q}\} \) is contained in the family \( \{\hat{q}(X) : \hat{q} \in \hat{\mathcal{Q}}\} \). Similarly, if \( \mathcal{M} \) is a convex family of measures then the family \( \{d\hat{m}/dP(X) : \hat{m} \in \hat{\mathcal{M}}\} \) is contained in the family \( \{d\mathcal{m}/dP(X) : \mathcal{m} \in \mathcal{M}\} \).

For finite \( \mathcal{X} = \{1, \ldots, J\} \), both the space of nonnegative functions \( q = [q_j]_{1 \leq j \leq J} \) on \( \mathcal{X} \) and the space of nonnegative measures \( m = [m_j]_{1 \leq j \leq J} \) on \( \mathcal{X} \) can be identified with \( \mathcal{R}_+^J \). Two convex corners of nonnegative functions and measures on \( \mathcal{X} \) are said to form an antiblocking pair if each is the polar of the other. Antiblocking pairs of convex polyhedra were introduced in combinatorics by Fulkerson [50] [51]. In particular, a pair of convex corners is associated with every finite undirected graph \( G \) by taking the convex hull of the indicator vectors of cliques and the convex hull of the indicator vectors of independent vertex sets. (A set of vertices of \( G \) is called a clique or an independent set if every two vertices in the set are adjacent or nonadjacent, respectively.) These convex corners are contained in the polar of each other because a clique and an independent set have at most one vertex in common, so the inner product of the corresponding indicator vectors is at most one. Csizszár et al. [37] proved that these convex corners form an antiblocking pair iff the graph \( G \) is perfect. These authors also defined the entropy of a convex corner \( \mathcal{Q} \) in \( \mathcal{R}_+^J \) as

\[
H_\mathcal{Q}(P) = \min_{q \in \mathcal{Q}} \sum_{1 \leq j \leq J} P_j \log \left( \frac{1}{q_j} \right).
\]

Obviously \( H_\mathcal{Q}(P) = -W_\mathcal{Q}(P) \) for any distribution \( P \) on \( \{1, \ldots, J\} \). If \( \mathcal{Q} \) is the convex corner of sub-probability measures on \( \{1, \ldots, J\} \) then the minimum is attained when \( q \) is equal to the true probability mass function \( P \), and the minimum \( H_\mathcal{Q}(P) \) is equal to the Shannon entropy \( H(P) = \sum_{1 \leq j \leq J} P_j \log(1/P_j) \). Many applications of convex corners in combinatorics and information theory are discussed in the survey paper of Simonyi [115].

**B. The minimax property**

We prove that the maximum growth exponent \( W_\mathcal{Q}(P) \) for a convex family \( \mathcal{Q} \) is equal to the minimum information functional \( I_{\mathcal{M}(\mathcal{Q})}(P) \) for the polar family \( \mathcal{M}(\mathcal{Q}) \) and vice versa.

**Theorem 15.** Let \( X \) be a random variable with distribution \( P \) on the space \( \mathcal{X} \).
(a) Let \( \mathcal{Q} \) be a convex family of functions on \( \mathcal{X} \) such that \( W_\mathcal{Q}(P) \) is well defined and finite. If \( q^* \) is log-optimum in \( \hat{\mathcal{Q}} \), then \( dm^* = dP/q^* \) is log-optimum in the polar family \( \mathcal{M}(\mathcal{Q}) \), \( q^* \) is log-optimum in the convex corner \( \hat{\mathcal{Q}} \), and the maximum growth exponent \( W_\mathcal{Q}(P) \) is equal to the minimum information \( I_{\mathcal{M}(\mathcal{Q})}(P) \). In fact, we have pointwise equality

\[
q^*(X) = \frac{dP}{dm^*}(X) \quad \text{almost surely.}
\]

(b) Dually, suppose \( \mathcal{M} \) is a convex family of measures on \( \mathcal{X} \) such that \( I_{\mathcal{M}(\mathcal{Q})}(P) \) is well defined and finite. If \( m^*(dx) \) is log-optimum in \( \hat{\mathcal{M}} \) then \( q^*(x) = dP/dm^*(x) \) is log-optimum in the polar family \( \mathcal{Q}(\mathcal{M}) \), \( m^*(dx) \) is log-optimum in the convex corner \( \hat{\mathcal{M}} \), and the minimum information \( I_{\mathcal{M}(\mathcal{Q})}(P) \) is equal to the maximum growth exponent \( W_{\mathcal{Q}(\mathcal{M})}(P) \).
Proof: If \( q^* \) is log-optimum in \( \mathcal{Q} \) then the measure \( d m^* = dP/q^* \) belongs to \( \mathcal{M}(\mathcal{Q}) \) since

\[
\int q \, dm^* = E_P \left\{ \frac{q(X)}{q^*(X)} \right\} \leq 1, \quad \text{for all } q \in \mathcal{Q}.
\]

Dually, if \( m^* \) in log-optimum in \( \mathcal{M} \) then the function \( q^* = dP/dm^* \) belongs to \( \mathcal{Q}(\mathcal{M}) \) since

\[
\int q^* \, dm = E_P \left\{ \frac{d m}{d m^*}(X) \right\} \leq 1, \quad \text{for all } m \in \mathcal{M}.
\]

If \( q^* = dP/dm^* \) or \( dm^* = dP/q^* \) then obviously

\[
W_{q^*}(P) = E_P \{ \log q^*(X) \} = E_P \left\{ \log \left( \frac{dP}{dm^*}(X) \right) \right\} = I_{m^*}(P).
\]

The theorem follows. \( \blacksquare \)

Let \( \mathcal{Q} \) and \( \mathcal{M} \) be convex families of random variables and measures on \( X \) such that \( \int q \, dm \leq 1 \) for all \( q \in \mathcal{Q} \) and \( m \in \mathcal{M} \). Choose \( q_k \in \mathcal{Q} \) such that the growth exponent of \( q_k \) converges to the maximum \( W_{\mathcal{Q}}(P) \), and choose \( m_k \in \mathcal{M} \) such that the information of \( P \) relative to \( m_k \) converges to the minimum \( I_{\mathcal{M}}(P) \). If \( W_{\mathcal{Q}}(P) \) and \( I_{\mathcal{M}}(P) \) are well defined and finite then by Theorem 7 and its dual there exist \( q^* \in \mathcal{Q} \) and \( m^* \in \mathcal{M} \) such that

\[
\log q_k(X) \to \log q^*(X) \quad \text{in } L^1(P),
\]

\[
\log \left( \frac{dP}{dm_k}(X) \right) \to \log \left( \frac{dP}{dm^*}(X) \right) \quad \text{in } L^1(P).
\]

If the inequality \( W_{\mathcal{Q}}(P) \leq I_{\mathcal{M}}(P) \) holds with equality then \( q^* = dP/dm^* \) almost surely and the subprobability \( q_k \, dm_k \) converges to \( dP = q^* \, dm^* \) in information since

\[
I_{q_k dm_k}(P) = E_P \left\{ \log \left( \frac{q^* dm^*}{q_k dm_k} \right) \right\} = [W_{\mathcal{Q}}(P) - E_P \{ \log q_k(X) \}] + [I_{m_k}(P) - I_{\mathcal{M}}(P)]
\]

is to sum of two nonnegative terms that converge to 0.

All these results generalize in a straightforward manner when the selections can be made with knowledge of side information \( \mathcal{G} \). The maximum growth exponent over a \( \mathcal{G} \)-convex family \( \Gamma \) of measurable functions \( \gamma(x|\mathcal{G}) \geq 0 \) is defined as

\[
W_{\Gamma}(X|\mathcal{G}) = \sup_{\gamma(x|\mathcal{G}) \in \Gamma} E\{ \log \gamma(X|\mathcal{G}) \}.
\]

Dually, the minimum \( \mathcal{G} \)-conditional information relative to a \( \mathcal{G} \)-convex family \( \Pi \) of transition kernels \( \mu(dx|\mathcal{G}) \) is defined as

\[
I_{\Pi}(X|\mathcal{G}) = \inf_{\mu(dx|\mathcal{G})} E \left\{ -\log \left( \frac{d\mu}{dP}(X|\mathcal{G}) \right) \right\}.
\]
Observe that $W_T(X|\mathcal{G}) = E\{w_T(X|\mathcal{G})\}$ and $I_\Pi(X|\mathcal{G}) = E\{i_\Pi(X|\mathcal{G})\}$ where

$$w_T(X|\mathcal{G}) = \text{ess sup}_{\gamma(x|\mathcal{G}) \in \Gamma} E\{\log \gamma(X|\mathcal{G})|\mathcal{G}\},$$

$$i_\Pi(X|\mathcal{G}) = \text{ess inf}_{\mu(dx|\mathcal{G})} E\left\{-\log \left(\frac{d\mu}{d\mathcal{P}}(X|\mathcal{G})\right)|\mathcal{G}\right\}.$$

If $\int \gamma(x|\mathcal{G}) \mu(dx|\mathcal{G}) \leq 1$ almost surely for all $\gamma(x|\mathcal{G})$ in $\Gamma$ and all $\mu(dx|\mathcal{G})$ in $\Pi$, then $w_T(X|\mathcal{G}) \leq i_\Pi(X|\mathcal{G})$ and consequently $W_T(X|\mathcal{G}) \leq I_\Pi(X|\mathcal{G})$. If $\Pi$ is the $\mathcal{G}$-conditional polar of $\Gamma$ or vice versa then $w_T(X|\mathcal{G}) = i_\Pi(X|\mathcal{G})$ almost surely and consequently

$$W_T(X|\mathcal{G}) = I_\mathcal{M}(X|\mathcal{G}).$$

The conditionally log-optimum $\gamma^*(x|\mathcal{G})$ in $\Gamma$ and $\mu^*(dx|\mathcal{G})$ in $\Pi$ then define the decomposition $P(dx|\mathcal{G}) = \gamma^*(x|\mathcal{G}) \mu^*(dx|\mathcal{G})$ and we have pointwise equality

$$\gamma^*(X|\mathcal{G}) = \frac{d\mathcal{P}}{d\mu^*}(X|\mathcal{G}) \text{ almost surely.}$$

C. Relativized growth exponents

As before let $X$ be a random variable with distribution $P$ on a measurable space $\mathcal{X}$ and let $Q$ be a convex family of nonnegative measurable functions on $\mathcal{X}$. We express the growth exponent of selections in $Q$ not in absolute terms but relative to a suitable reference. As reference we take a function $\gamma(x)$ such that $\gamma(X) > 0$ almost surely. Note that $q^*$ is log-optimum in $Q$ iff $q^*/\gamma$ is log-optimum in the relativized family

$$Q/\gamma = \{q/\gamma : q \in Q\}.$$

The maximum growth exponent of selections in $Q$ relative to $\gamma$ is defined as the maximum growth exponent of selections in the relativized family $Q/\gamma$:

$$W_{Q/\gamma}(P) = \sup_{q \in Q} E_P \left\{\log \left(\frac{q(X)}{\gamma(X)}\right)\right\}.$$ 

The maximum growth exponent $W_Q(P)$ is the sum of the growth exponent $W_{\gamma}(P) = E_P\{\log \gamma(X)\}$ and the maximum growth exponent relative to $\gamma$:

$$W_Q(P) = W_{\gamma}(P) + W_{Q/\gamma}(P).$$

Recall that $W_{Q/\gamma}(P) = I_{\mathcal{M}(Q/\gamma)}(P)$. We now assume that $\gamma \in Q$ so that $W_{Q/\gamma}(P) \geq 0$.

Lemma 4. The convex corner $\mathcal{M}(Q/\gamma)$ is equal to $\gamma \mathcal{M}(Q)$, and if $\gamma \in Q$ then it contains only subprobability measures. If $q^*$ is log-optimum in $Q$ then $dm^* = dP/q^*$ is log-optimum in $\mathcal{M}(Q)$, $dM^* = \gamma dm^*$ is log-optimum in the convex corner $\gamma \mathcal{M}(Q) = \mathcal{M}(Q/\gamma)$, and

$$W_{Q/\gamma}(P) = I_{\mathcal{M}^*}(P) = I_{\mathcal{M}(Q/\gamma)}(P).$$
Proof: Any measure \( m \in \mathcal{M}(Q) \) defines a measure \( dM = \gamma dm \) in \( \mathcal{M}(Q/\gamma) \) since
\[
\int \frac{q}{\gamma} dM = \int \frac{q}{\gamma} \gamma dm \leq \int q dm \leq 1, \quad \text{for all } q \in Q.
\]
Conversely, if \( M \in \mathcal{M}(Q/\gamma) \) then \( dm = dM/\gamma \) is a measure in \( \mathcal{M}(Q) \) since
\[
\int q dm = \int \frac{q}{\gamma} dM \leq 1, \quad \text{for all } q \in Q.
\]
Thus \( \mathcal{M}(Q/\gamma) = \gamma \mathcal{M}(Q) \). If \( \gamma \in \Gamma \) then every measure \( M \in \mathcal{M}(Q/\gamma) \) is subnormalized since \( \int (q/\gamma) dM \leq 1 \) for all \( q \in Q \) and in particular \( M(\mathcal{X}) = \int (\gamma/\gamma) dM \leq 1 \). The rest is obvious.

The normalized measures \( M \in \gamma \mathcal{M}(Q) \) or \( \mathcal{M}(Q/\gamma) \) are exactly the probability distributions under which \( \gamma \) is log-optimum in \( Q \). Indeed, the conditions for membership of \( M \) in \( \mathcal{M}(Q/\gamma) \) are precisely the KKT conditions for log-optimality of \( \gamma \) in \( Q \) under distribution \( M \). Notice that the set of distributions \( M \) for which \( \gamma \) is log-optimum in \( Q \) is nonempty iff \( \gamma \) is exposed in \( Q \) in the sense that \( \int \gamma dm = 1 \) for some \( m \in \mathcal{M}(Q) \). Indeed, if \( \gamma \) is log-optimum in \( Q \) under distribution \( M \) then \( dm = dM/\gamma \) is a measure in \( \mathcal{M}(Q) \) such that \( \int \gamma dm = 1 \), and conversely if \( \int \gamma dm = 1 \) for some \( m \in \mathcal{M}(Q) \) then \( \gamma \) is log-optimum under the distribution \( dM = \gamma dm \).

**Theorem 16.** Let \( X \) be a random variable with distribution \( P \) and let \( Q \) be a convex family of nonnegative measurable functions on the space \( \mathcal{X} \). Suppose \( \gamma \in Q \), \( q^* \) is log-optimum in \( Q \), and the KKT condition \( E_P\{\gamma/q^*\} \leq 1 \) holds with equality. Then the maximum relativized growth exponent \( W_{Q/\gamma}(P) \) is equal to the minimum information divergence of \( P \) relative to probability distributions on \( \mathcal{X} \) under which \( \gamma \) is log-optimum in \( Q \). The minimum divergence is attained relative to the probability distribution \( dM^* = \gamma dm^* = (\gamma/q^*) dP \).

Theorem 16 can be proved using Stein’s lemma. In fact, \( M^* \) is the worst alternative when discriminating the null hypothesis \( P \) against alternative distributions in the convex family \( \mathcal{M}(Q/\gamma) \). The likelihood ratio test of \( P \) against \( M^* \) based on independent realizations \( X_i \) of \( X \) is nothing but the capital ratio test which rejects \( P \) when \( \prod_{i=1}^n q^*(X_i)/\gamma(X_i) \) does not grow exponentially fast with a positive rate. The probability of rejecting \( P \) decreases exponentially with rate \( I_{M^*}(P) = W_{Q/\gamma}(P) \).

Stein’s lemma was first used by Móri [92] to prove a special case of Theorem 16 for log-optimum investment. Móri proved that the maximum capital growth rate \( W_Q(P) \) is equal to the minimum information divergence \( I_{\mathcal{M}(Q)}(P) \) of the true market distribution \( P \) relative to unfavorable market distributions for which saving all capital in the form of cash is a log-optimum strategy. This essentially is Theorem 16 in the special case when \( \gamma(x) \equiv 1 \). We shall prove a more general result using a different method.

The loss in achievable growth exponent if we are no longer free to select in the convex family \( Q \) but are constrained to some convex subfamily \( \Gamma \subseteq Q \) will be denoted by

\[
W_{Q/\Gamma}(P) = \sup_{q \in Q} \inf_{\gamma \in \Gamma} \left\{ \log \left( \frac{q(X)}{\gamma(X)} \right) \right\}.
\]

The probability measures \( M \) under which the log-optimum selection in \( Q \) happens to fall in \( \Gamma \) and hence \( W_{Q/\Gamma}(M) = 0 \) are exactly the normalized distributions in \( \Gamma \mathcal{M}(Q) \). We prove that \( W_{Q/\Gamma}(P) \) is the minimum information divergence of \( P \) relative to such probability
measures \( M \) for which the constraint \( \Gamma \) is inconsequential in terms of achievable growth exponent. The previous theorem follows by specializing \( \Gamma \) to the singleton set \( \{ \gamma \} \).

**Theorem 17.** If \( \Gamma \) is a convex subfamily of the convex family \( \mathcal{Q} \), then the maximum relativized growth exponent \( W_{\mathcal{Q}/\Gamma}(P) \) is equal to the minimum information divergence of \( P \) relative to subprobability measures \( \gamma \), \( dm \) in the (generally non-convex) family \( \Gamma \mathcal{M}(\mathcal{Q}) \):

\[
W_{\mathcal{Q}/\Gamma}(P) = I_{\Gamma \mathcal{M}(\mathcal{Q})}(P).
\]

If \( q^* \) is log-optimum in \( \mathcal{Q} \) and \( \gamma^* \) is log-optimum in \( \Gamma \) then the minimum divergence is attained relative to the subprobability measure

\[
dM^* = \gamma^* dm^* = \frac{\gamma^*}{q^*} dP.
\]

If \( M^* \) is normalized (e.g. if \( q^* \) falls in the algebraic interior of \( \mathcal{Q} \)), then \( \gamma^* \) is log-optimum in \( \mathcal{Q} \) under \( M^* \) and \( W_{\mathcal{Q}/\Gamma}(P) \) is equal to the minimum information divergence of \( P \) relative to probability measures under which the log-optimum selection in \( \mathcal{Q} \) happens to fall in \( \Gamma \).

**Proof.** If \( q^* \) is log-optimum in \( \mathcal{Q} \) and \( \gamma^* \) is log-optimum in \( \Gamma \) then the subprobability measure \( dM^* = \gamma^* dm^* = (\gamma^*/q^*) dP \) is log-optimum in \( \Gamma \mathcal{M}(\mathcal{Q}) \) and

\[
W_{\mathcal{Q}/\Gamma}(P) = E_P \left\{ \log \left( \frac{q^*(X)}{\gamma^*(X)} \right) \right\} = I_{M^*}(P) = I_{\Gamma \mathcal{M}(\mathcal{Q})}(P).
\]

If \( q^* \) falls in the algebraic interior of \( \mathcal{Q} \) then \( E_P \{\gamma^*/q^*\} = 1 \) and \( dM^* = (\gamma^*/q^*) dP \) is a probability measure. The rest of the proof is easy. \( \blacksquare \)

As always, there are parallels between this result for maximum growth exponents and the dual counterpart for minimum information rates. Suppose \( \mathcal{M} \) and \( \Pi \) are convex families of \( \sigma \)-finite measures on \( X \). If \( \mathcal{M} \subseteq \Pi \) then one may define

\[
I_{\mathcal{M} \setminus \Pi}(P) = \inf_{m \in \mathcal{M}} \sup_{\mu \in \Pi} E_P \left\{ \log \left( \frac{d\mu}{dm}(X) \right) \right\}.
\]

If \( I_{\mathcal{M}}(P) \) and \( I_{\Pi}(P) \) are well defined and finite then

\[
I_{\mathcal{M} \setminus \Pi}(P) = I_{\mathcal{M}}(P) - I_{\Pi}(P) \geq 0.
\]

Suppose \( \mu^* \) and \( \mu^* \) are log-optimum in \( \mathcal{M} \) and \( \Pi \), respectively. The KKT condition \( E_P \left\{ \frac{d\mu^*}{dm^*}(X) \right\} \leq 1 \) holds with equality if the subprobability measure \( dM^* = dP \, dm^*/d\mu^* \) is normalized. If this is the case, then \( I_{\mathcal{M} \setminus \Pi}(P) \) is equal to the minimum information divergence rate of \( P \) relative to probability distributions under which the log-optimum \( m^* \) in \( \mathcal{M} \) happens to be log-optimum in the larger family \( \Pi \).

Suppose \( \Pi = \mathcal{M}(\Gamma) \) and \( \mathcal{M} = \mathcal{M}(\mathcal{Q}) \) are the polars of convex corners \( \Gamma \) and \( \mathcal{Q} \). If \( \Gamma \subseteq \mathcal{Q} \), then \( \Pi \supseteq \mathcal{M} \) and we have equality \( W_{\mathcal{Q}/\Gamma}(P) = I_{\mathcal{M} \setminus \Pi}(P) \). In fact, suppose \( q^*(x) \) and \( \gamma^*(x) \) are log-optimum in \( \mathcal{Q} \) and \( \Gamma \), and \( m^*(dx) \) and \( \mu^*(dx) \) are log-optimum in the polar families \( \mathcal{M} \) and \( \Pi \). Since \( P(dx) = q^*(x) m^*(dx) \) and \( P(dx) = \gamma^*(x) \mu^*(dx) \), we have \( q^*(X)/\gamma^*(X) = d\mu^*/dm^*(X) \). Clearly, \( W_{\mathcal{Q}/\Gamma}(P) = I_{\mathcal{M} \setminus \Pi}(P) \) is equal to the divergence \( I_{M^*}(P) \) where

\[
M^*(dx) = \gamma^*(x) m^*(dx) = \frac{\gamma^*(x)}{q^*(x)} P(dx) = \frac{\mu^*(dx)}{m^*(dx)} P(dx).
\]
D. The convex conjugate of $I_\mathcal{M}(P)$ and $W_\mathcal{Q}(P)$

Let $\mathcal{M}$ be a convex set of σ-finite measures on $\mathcal{X}$. The functional $I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} I_m(P)$ is convex in $P$ because $I_\mathcal{M}(P) = W_{\mathcal{Q}(\mathcal{M})}(P)$ and the functional $W_{\mathcal{Q}(\mathcal{M})}(P)$ is convex. The epigraph of $I_\mathcal{M}(P)$ is the projection of the epigraph of the functional $I_m(P)$, which is convex jointly in $P$ and $m \in \mathcal{M}$. If the total mass $m(\mathcal{X})$ of every measure $m \in \mathcal{M}$ is bounded by a constant $c$, then $\mathcal{Q}(\mathcal{M})$ contains the constant function $\gamma(x) \equiv 1/c$. By Theorem 9, $I_\mathcal{M}(P) = W_{\mathcal{Q}(\mathcal{M})}(P)$ is bounded below (by $-\log c$) and lower semicontinuous in $P$ for the weak or the strong topology. In particular, $I_\mathcal{M}(P)$ is nonnegative convex and lower semicontinuous in $P$ if $\mathcal{M}$ is a convex family of probability or sub-probability measures.

Recall that the convex conjugate of the functional $I_m(P)$ is equal to $\Lambda_m(V) = \log(\int e^V dm)$. We describe the convex conjugate of $I_\mathcal{M}(P)$ for any convex family $\mathcal{M}$.

**Theorem 18.** If $\mathcal{M}$ is a convex family of σ-finite measures on $\mathcal{X}$, then the minimum information functional $I_\mathcal{M}(P) = \inf_{m \in \mathcal{M}} I_m(P)$ is convex in $P$ with convex conjugate

$$\Lambda_\mathcal{M}(V) = \sup_{m \in \mathcal{M}} \Lambda_m(V). \quad (1.2)$$

**Proof:** Observe that

$$\sup_P [(P, V) - I_\mathcal{M}(P)] = \sup_P [(P, V) - \inf_{m \in \mathcal{M}} I_m(P)]$$

$$= \sup_P \sup_{m \in \mathcal{M}} [(P, V) - I_m(P)]$$

$$= \sup_{m \in \mathcal{M}} \sup_P [(P, V) - I_m(P)]$$

$$= \sup_{m \in \mathcal{M}} \Lambda_m(V) = \Lambda_\mathcal{M}(V).$$

Thus the convex conjugate functional of $I_\mathcal{M}(P)$ is equal to $\Lambda_\mathcal{M}(V)$. □

If $Q$ is a convex family of nonnegative measurable functions on $\mathcal{X}$ then $W_{\mathcal{Q}}(P) = I_{\mathcal{M}(\mathcal{Q})}(P)$, so the convex conjugate of $W_{\mathcal{Q}}(P)$ is equal to $\Lambda_{\mathcal{M}(\mathcal{Q})}(V)$. If $\gamma \in \mathcal{Q}$ then the graph of the affine functional $W_\gamma(P) = E_P\{\log \gamma(X)\}$ is a hyperplane below the graph of the convex functional $W_{\mathcal{Q}}(P)$. If we raise this hyperplane by the amount $-\Lambda_{\mathcal{M}(\mathcal{Q})}(V)$ where $V(x) = \log \gamma(x)$, then it supports the graph of $W_{\mathcal{Q}}(P)$. Suppose the supremum $\Lambda_{\mathcal{M}(\mathcal{Q})}(V) = \sup_{m \in \mathcal{M}(\mathcal{Q})} \Lambda_m(V)$ is finite and attained by $m \in \mathcal{M}(\mathcal{Q})$. Then

$$\Lambda_{\mathcal{M}(\mathcal{Q})}(V) = \log(\int e^V dm) = \log(\int \gamma dm) \leq 0, \quad (1.3)$$

and the supporting hyperplane with slope $V = \log \gamma$ will support the graph of $W_{\mathcal{Q}}(P)$ at the contact point $P_\gamma$ where

$$dP_\gamma = \frac{e^V dm}{\int e^V dm} = \frac{\gamma dm}{\int \gamma dm}. \quad (1.4)$$

The function $p_\gamma = dP_\gamma/dm = \gamma/(\int \gamma dm)$ is log-optimum in the convex corner $Q$ and the measure $m(dx)$ is log-optimum in the polar family $\mathcal{M}(Q)$ under distribution $dP_\gamma = p_\gamma dm$.

The relativized growth exponent $W_{\mathcal{Q}/\gamma}(P) = W_{\mathcal{Q}}(P) - W_\gamma(P)$ is the vertical distance at $P$ between the epigraph of the convex functional $W_{\mathcal{Q}}(P)$ and the hyperplane with slope $\log \gamma$ that is the graph of the linear functional $W_\gamma(P) = E_P\{\log \gamma(X)\}$. If $\gamma$ is log-optimum in $Q$ under some distribution $M(dx)$, then that hyperplane supports the epigraph at $M$.
and the distance $W_{Q/\gamma}(P)$ vanishes when $P = M$. Thus $W_\gamma(P) = W_\gamma(P) + W_{Q/\gamma}(P)$ is the sum of the linear functional $W_\gamma(P)$ and the nonnegative convex functional $W_{Q/\gamma}(P)$ that vanishes at $M$. Therefore one may regard $\log \gamma$ as the gradient and $W_{Q/\gamma}(P)$ as the second order remainder term in the Taylor series expansion of the convex functional $W_Q(P)$ at $M$.

**E. Support and gauge functionals**

Associated with every family $Q$ of measurable functions $q(x) \geq 0$ on $\mathcal{X}$ is a support functional which uniquely determines the polar family $\mathcal{M}(Q)$ and the convex corner $\hat{Q}$. The support functional of $Q$ is defined for every measure $m(dx)$ on $\mathcal{X}$ as

$$\sigma_Q(m) = \sup_{q \in Q} \int q \, dm.$$ 

Clearly, $m \in \mathcal{M}(Q)$ iff $\sigma_Q(m) \leq 1$. Dually, if $\mathcal{M}$ is a set of measures $m(dx)$ on $\mathcal{X}$ then the support functional of $\mathcal{M}$ is defined for every measurable function $q(x) \geq 0$ as

$$\sigma_M(q) = \sup_{m \in \mathcal{M}} \int q \, dm.$$ 

The polar family $\mathcal{Q}(\mathcal{M})$ is the set of measurable functions $q(x) \geq 0$ such that $\sigma_M(q) \leq 1$. Support functionals are nonnegative, convex, positively homogeneous, and monotone in the sense that $\sigma_Q(m') \leq \sigma_Q(m)$ if $0 \leq m' \leq m$ and $\sigma_M(q') \leq \sigma_M(q)$ if $0 \leq q' \leq q$. Two families with the same support functional have the same polar and therefore they define the same convex corner.

The gauge or Minkowski functional of $Q$ is defined for measurable $q(x) \geq 0$ as

$$\rho_Q(q) = \inf\{r > 0 : r^{-1}q \in Q\}.$$ 

By convention, $\rho_Q(q) = \infty$ when the infimum is taken over the empty set. If $Q$ is down-monotone and convex and $Q$ generates the convex corner $\hat{Q}$, then

$$\{q \mid q \geq 0, \rho_Q(q) < 1\} \subseteq Q \subseteq \{q \mid q \geq 0, \rho_Q(q) \leq 1\} \subseteq \hat{Q}.$$ 

The maximum growth exponent of $Q$ is then the same as that of $\{q \mid q \geq 0, \rho_Q(q) \leq 1\}$. The functional $\rho_Q(q) = \rho_Q(|q|)$ has the properties of a norm, except that $\rho_Q(q) = 0$ and $\rho_Q(q) = \infty$ are permissible values. If $\rho_Q$ happens to be a norm indeed, then $\{q \mid q \geq 0, \rho_Q(q) < 1\}$ and $\{q \mid q \geq 0, \rho_Q(q) \leq 1\}$ are the intersections of the positive cone with the open and closed unit balls for the norm $\rho_Q$.

If $\mathcal{M}$ is a convex down-monotone family of measures on $Q$ then the gauge or Minkowski functional of $\mathcal{M}$ is defined for arbitrary measures $m(dx)$ on $\mathcal{X}$ as

$$\rho_M(m) = \inf\{r > 0 : r^{-1}m \in \mathcal{M}\}.$$ 

It is easy to verify that the support functional of a convex corner is related to the gauge functional of the polar convex corner as follows:

$$\sigma_Q(m) \leq \rho_M(Q)(m), \quad \sigma_M(g) \leq \rho_Q(M)(q).$$ 

Thus the following chains hold with equality throughout:

$$Q(M) \subseteq \{q : \rho_M(M)(q) \leq 1\} \subseteq \{q : \sigma_M(q) \leq 1\} = Q(M),$$

45
\[ \mathcal{M}(\mathcal{Q}) \subseteq \{ m : \rho_{\mathcal{M}(\mathcal{Q})}(m) \leq 1 \} \subseteq \{ m : \sigma_{\mathcal{Q}}(m) \leq 1 \} = \mathcal{M}(\mathcal{Q}). \]

If \( \mathcal{Q} \) and \( \mathcal{M} \) are polars of each other then for any \( q(x) \) in \( \mathcal{Q} \) such that \( \sigma_{\mathcal{M}}(q) < \infty \) and any \( m(dx) \) in \( \mathcal{M} \) such that \( \rho_{\mathcal{M}}(m) < \infty \) we have Hölder's inequality

\[ \int q \, dm \leq \sigma_{\mathcal{M}}(q) \rho_{\mathcal{M}}(m) \leq \rho_{\mathcal{Q}}(q) \rho_{\mathcal{M}}(m). \]

Indeed, the normalized function \( \hat{q} = q/\sigma_{\mathcal{M}}(q) \) is contained in the convex corner \( \hat{\mathcal{Q}} \) and hence \( \int \hat{q} \, dm \leq r \) if \( r^{-1}m \in \mathcal{M} \). Similarly, if \( \rho_{\mathcal{Q}}(q) < \infty \) and \( \sigma_{\mathcal{Q}}(m) < \infty \) then

\[ \int q \, dm \leq \rho_{\mathcal{Q}}(q) \sigma_{\mathcal{Q}}(m) \leq \rho_{\mathcal{Q}}(q) \rho_{\mathcal{M}}(m). \]

**F. Function seminorms**

A natural framework for studying convex corners is the theory of Banach function spaces. Concrete examples of these are Lebesgue's \( L_p \) spaces, Orlicz spaces, and the perfect sequence spaces (volkommene Räume) of Köthe and Toeplitz [75]. For a complete treatment see Chapter 15 of Zaanen [129] and the comprehensive treatise of Zaanen [130].

Normed function spaces are spaces of measurable functions on a complete \( \sigma \)-finite measure space \( (\mathcal{X}, \lambda) \). We really deal with equivalence classes modulo functions that vanish \( \lambda \)-almost everywhere. A functional \( \rho(q) \) which is defined for all measurable functions \( q(x) \geq 0 \) on \( \mathcal{X} \) is called a function seminorm if the following conditions are satisfied.

(i). \( 0 \leq \rho(q) \leq \infty \) (the value \( +\infty \) is allowed);

(ii). \( \rho(aq) = a\rho(q) \) for any constant \( a > 0 \);

(iii). \( \rho(q + q') \leq \rho(q) + \rho(q') \);

(iv). if \( 0 \leq q' \leq q \) then \( \rho(q') \leq \rho(q) \).

For any real (or, complex) valued measurable function \( q(x) \) on \( \mathcal{X} \) let \( \rho(q) = \rho(|q|) \), and let \( L_{\rho} \) denote the linear space of all measurable functions \( q(x) \) such that \( \rho(q) < \infty \). If \( q(x) \) and \( q'(x) \) are measurable functions such that \( |q'(x)| \leq |q(x)| \) and \( q \in L_{\rho} \), then also \( q' \in L_{\rho} \) and \( \rho(q') \leq \rho(q) \). Thus the functional \( \rho \) is a seminorm on \( L_{\rho} \), and this seminorm is monotone and absolute. If \( \rho \) is a norm then \( L_{\rho} \) is called a normed Köthe space or, when it is norm complete, a Banach function space. The intersection of the positive cone with the unit ball of \( L_{\rho} \) is convex and down-monotone, and will be denoted by

\[ \mathcal{Q}_{\rho} = \{ q | q \geq 0, \rho(q) \leq 1 \}. \]

A function seminorm \( \rho \) is said to have the Fatou property if for arbitrary measurable functions \( q_k(x) \) and \( q(x) \) such that \( 0 \leq q_k(x) \searrow q(x) \), we have \( \rho(q_k) \searrow \rho(q) \). For this, it is necessary and sufficient that for any measurable functions \( q_n(x) \geq 0, \)

\[ \rho(\lim \inf_k q_k) \leq \lim \inf_k \rho(q_k). \]

For any function seminorm \( \rho \), there exists a largest function seminorm \( \hat{\rho} \) such that \( \hat{\rho} \leq \rho \) and \( \hat{\rho} \) has the Fatou property. It is defined for measurable \( q(x) \geq 0 \) as

\[ \hat{\rho}(q) = \inf_{0 \leq q_k \searrow q} \lim_k \rho(q_k). \]

In particular, \( \rho \) has the Fatou property iff \( \rho = \hat{\rho} \). If \( \{ \rho_i \} \) is a class of function seminorms with the Fatou property then also \( \sup \rho_i \) is a function seminorm with the Fatou property. Thus the class of function seminorms with the Fatou property is a complete lattice.
The associate seminorm $\rho'(m)$ is defined for any measurable function $m(x)$ on $\mathcal{X}$ as

$$\rho'(m) = \sup_{q : \rho(q) \leq 1} \int |q(x) m(x)| \lambda(dx).$$

The space $L'_\rho$ of measurable functions $m(x)$ such that $\rho'(m) < \infty$ is a Banach function space with seminorm $\rho'$, and is called the associate space of $L_\rho$. For arbitrary measurable functions $q(x)$ and $m(x)$ such that both $\rho(q)$ and $\rho'(m)$ are finite we have Hölder's inequality

$$\left| \int q(x) m(x) \lambda(dx) \right| \leq \int |q(x) m(x)| \lambda(dx) \leq \rho(q) \rho'(m).$$

The associate function seminorm $\rho'(m)$ has the Fatou property, and the second associate $\rho''(q)$ is equal to the largest function seminorm with the Fatou property below $\rho(q)$:

$$\rho''(q) = \hat{\rho}(q).$$

Third and higher associate function seminorms give nothing new since $\rho'''(m) = \rho'(m)$. Indeed, if $\rho_1 \leq \rho_2$ then $\rho'_1 \geq \rho'_2$. Since $\rho'' \leq \rho$, we see that $\rho' \leq (\rho'')' = (\rho')'' \leq \rho'$.

The positive unit ball of $L'_\rho$ is denoted by

$$\mathcal{M}_\rho = \mathcal{Q}_{\rho'} = \{ m \mid m \geq 0, \rho'(m) \leq 1 \}.$$

Any measurable function $m(x) \geq 0$ can be regarded as the density of a $\sigma$-finite measure $m(dx) = m(x) \lambda(dx)$ on $\mathcal{X}$, and $\mathcal{M}_\rho$ may be identified with the set of $\lambda$-dominated measures in the polar $\mathcal{M}(\mathcal{Q}_\rho)$. The polar of $\mathcal{M}_\rho$ is equal to the positive unit ball $\mathcal{Q}_\rho = Q_\rho = Q_{\rho''}$ for the second associate function seminorm $\hat{\rho} = \rho''$. This is the convex corner generated by $Q_\rho$. If $\rho$ has the Fatou property then $\rho = \hat{\rho}$, and $\mathcal{Q}_\rho = \mathcal{Q}_\rho$ is a convex corner to begin with. Thus every convex corner is the intersection of the positive cone with the closed unit ball of a function seminorm with the Fatou property.

Suppose $\rho$ satisfies the Fatou property, so that $\mathcal{Q}_\rho$ and $\mathcal{M}_\rho$ are convex corners and polars of each other. It is shown in Theorem 6 of Lozanovskii [89] that the density $p(x)$ of every $\lambda$-dominated probability measure $P(dx) = p(x) \lambda(dx)$ can be written as the product

$$p(x) = q^*(x) m^*(x)$$

of a function $q^*(x)$ in $\mathcal{Q}_\rho$ and a function $m^*(x)$ in $\mathcal{M}_\rho$. This factorization property has been rediscovered by several authors including Jamison and Ruckle [66], Gillespie [56], and Saint-Raymond [110]. See also Reisner [106], Csizsár et al. [37], Bollobás and Leader [19].

The factorization $p(x) = q^*(x) m^*(x)$ defines a splitting $P(dx) = q^*(x) m^*(dx)$ where $q^*(x)$ and $m^*(dx) = m^*(x) \lambda(dx)$ are log-optimum in the convex corner $\mathcal{Q}_\rho$ and its polar $\mathcal{M}(\mathcal{Q}_\rho)$.

The best known examples of Banach function spaces are the Lebesgue $L_p$ spaces. If $1 \leq p \leq \infty$ then $L_p(\lambda)$ is a Banach function space with function norm

$$\rho_p(q) = \left[ \int |q(x)|^p \lambda(dx) \right]^{1/p}.$$
The theory of $L_p$ spaces has been generalized to the theory of Orlicz spaces. Let $\Phi(u)$ be a lower semicontinuous convex nondecreasing function of $u \geq 0$ such that $\Phi(0) = 0$. The Orlicz space $L_\Phi(\lambda)$ is defined as the linear space of measurable functions $q(x)$ such that $M_\Phi(k^{-1}q) < \infty$ for some $k > 0$, where

$$M_\Phi(q) = \int \Phi(|q|) \, d\lambda.$$  

It is well known that $L_\Phi(\lambda)$ is a Banach function space with the Luxemburg norm

$$\rho_\Phi(q) = \inf\{r \mid r > 0, M_\Phi(r^{-1}q) \leq 1\}.$$  

If $\rho_\Phi(q) \leq 1$ then $M_\Phi(q) \leq \rho_\Phi(q) \leq 1$ and if $\rho_\Phi(q) > 1$ then $M_\Phi(q) \geq \rho_\Phi(q) > 1$. Thus

$$\rho_\Phi(q) = \inf\{r \mid r > 0, r^{-1}q \in B_\Phi\}$$

is nothing but the gauge or Minkowski functional of the norm closed unit ball

$$B_\Phi = \{q \mid \rho_\Phi(q) \leq 1\} = \{q \mid M_\Phi(q) \leq 1\}.$$  

The Luxemburg norm $\rho_\Phi$ on $L_\Phi(\lambda)$ is a function seminorm with the Fatou property. The associate function seminorm $\rho_\Psi$ on the associate Banach function space $L'_\Phi(\lambda)$ is called the Orlicz norm on $L'_\Phi(\lambda)$. It turns out that $L'_\Phi(\lambda)$ is equal to the Orlicz space $L_\Psi(\lambda)$ where $\Psi$ is the Young conjugate of $\Phi$:

$$\Psi(v) = \sup_{u \geq 0} [uv - \Phi(u)], \quad v \geq 0.$$  

The Orlicz norm $\rho_\Phi$ is in general not equal to the Luxemburg norm $\rho_\Psi$ on $L_\Phi = L_\Psi$, but these norms are equivalent (precisely, $\rho_\Psi \leq \rho_\Phi \leq 2\rho_\Psi$). The unit balls $Q_\Phi = \{q \mid q \geq 0, \rho_\Phi(q) \leq 1\}$ in $L_\Phi$ and $M_\Phi = \{m \mid m \geq 0, \rho_\Phi(m) \leq 1\}$ in $L'_\Phi = L_\Psi$ are convex corners and polars of each other. The roles of $\Phi(u)$ and $\Psi(v)$ may be interchanged since $\Phi(u)$ is the Young conjugate of its Young conjugate $\Psi(v)$. Thus the unit ball $Q_\Psi$ for the Luxemburg norm $\rho_\Psi$ on $L_\Psi$ and the unit ball $M_\Psi$ for the Orlicz norm $\rho_\Psi$ on $L'_\Psi = L_\Phi$ are convex corners and polars of each other.

G. Interpolation between convex families

Let $Q_0$ and $Q_1$ be convex down-monotone families of nonnegative measurable functions on the space $\mathcal{X}$. It is possible to connect $Q_0$ and $Q_1$ by a scale of intermediate convex down-monotone families $Q_\theta$, $0 < \theta < 1$. Indeed, let $Q_\theta$ be defined as the down-monotone hull of $Q_0^{1-\theta}Q_1^\theta$. Thus $Q_\theta$ is the family of functions $q(x)$ such that $0 \leq q(x) \leq q_0(x)^{1-\theta}q_1(x)^\theta$ for some $q_0(x)$ in $Q_0$ and some $q_1(x)$ in $Q_1$. The family $Q_\theta$ is down-monotone and convex because both $Q_0$, $Q_1$ are down-monotone and convex and the function $F_\theta(q_0, q_1) = q_0^{1-\theta}q_1^\theta$ is concave in $(q_0, q_1)$. The gauge or Minkowski functional $\rho_\theta(q)$ of $Q_\theta$ is defined for any nonnegative measurable function $q(x)$ as the infimum of all values $r > 0$ such that the inequality $q(x) \leq r q_0(x)^{1-\theta}q_1(x)^\theta$ holds for some $q_0(x)$ in $Q_0$ and some $q_1(x)$ in $Q_1$. (By convention, $\rho_\theta(q) = \infty$ when the infimum is taken over the empty set.)

The closure $\bar{Q}_\theta$ of the interpolating family $Q_\theta$ contains the family $(\bar{Q}_0)^{1-\theta}\bar{Q}_1^\theta$ that interpolates between the closures of $Q_0$ and $Q_1$. The following result is also obvious.
Theorem 19. Let $X$ be a random variable with distribution $P$ on the space $\mathcal{X}$. If $q_0^\ast(x)$ and $q_1^\ast(x)$ are log-optimum in $\mathcal{Q}_0$ and $\mathcal{Q}_1$ then the log-optimum $q_\theta^\ast(x)$ in $\mathcal{Q}_\theta$ is given by

$$q_\theta^\ast(x) = q_0^\ast(x)^{1-\theta}q_1^\ast(x)^\theta.$$ 

It follows that the maximum growth exponent of $\mathcal{Q}_\theta$ varies linearly between the maximum growth exponents of $\mathcal{Q}_0$ and $\mathcal{Q}_1$ as $\theta$ varies between 0 and 1:

$$W_{\mathcal{Q}_\theta}(P) = (1 - \theta) W_{\mathcal{Q}_0}(P) + \theta W_{\mathcal{Q}_1}(P).$$

One may similarly connect two convex down-monotone families $\mathcal{M}_0$ and $\mathcal{M}_1$ of measures on $\mathcal{X}$ by a scale of intermediate convex down-monotone families $\mathcal{M}_\theta$, $0 < \theta < 1$. The family $\mathcal{M}_\theta$ is defined as the down-monotone hull of $\mathcal{M}_0^{1-\theta}\mathcal{M}_1^\theta$, that is the family of measures $m(dx)$ such that $m(dx) \leq m_0(dx)^{1-\theta}m_1(dx)^\theta$ for some $m_0(dx)$ in $\mathcal{M}_0$ and some $m_1(dx)$ in $\mathcal{M}_1$. Clearly $\mathcal{M}_\theta \supseteq (\mathcal{M}_0^{1-\theta}\mathcal{M}_1^\theta)$ and $m_\theta^\ast(dx) = m_0^\ast(dx)^{1-\theta}m_1^\ast(dx)^\theta$ is log-optimum in $\mathcal{M}_\theta$ if $m_0^\ast(dx)$ is log-optimum in $\mathcal{M}_0$ and $m_1^\ast(dx)$ is log-optimum in $\mathcal{M}_1$. Thus the minimum information relative to measures in $\mathcal{M}_\theta$ varies linearly as $\theta$ varies between 0 and 1:

$$I_{\mathcal{M}_\theta}(P) = (1 - \theta) I_{\mathcal{M}_0}(P) + \theta I_{\mathcal{M}_1}(P).$$

If $\mathcal{M}_0$ and $\mathcal{M}_1$ are the polars of convex families $\mathcal{Q}_0$ and $\mathcal{Q}_1$, then the convex down-monotone family $\mathcal{M}_\theta$ is contained in the polar of $\mathcal{Q}_\theta$. Indeed, let $q_0(x)$ and $q_1(x)$ be nonnegative measurable functions in $\mathcal{Q}_0$ and $\mathcal{Q}_1$ and let $m_0(dx)$ and $m_1(dx)$ be measures in the polar families $\mathcal{M}_0$ and $\mathcal{M}_1$. Then $q_\theta(x) = q_0(x)^{1-\theta}q_1(x)^\theta$ is a function in $\mathcal{Q}_\theta$, $m_\theta(dx) = m_0(dx)^{1-\theta}m_1(dx)^\theta$ is a measure in $\mathcal{M}_\theta$, and by Hölder's inequality

$$\int q_\theta dm_\theta \leq \left(\int q_0 dm_0\right)^{1-\theta}\left(\int q_1 dm_1\right)^\theta.$$ 

Clearly $\int q_\theta dm_\theta \leq 1$ because $\int q_i dm_i \leq 1$ for $i = 0, 1$. The argument that $\mathcal{M}_\theta$ is actually equal to the polar of $\mathcal{Q}_\theta$ is more delicate. The proof apparently requires introduction of locally convex topologies on dual spaces and invocation of the Hahn-Banach theorem.

The construction of the interpolating convex down-monotone families $\mathcal{Q}_\theta$ and $\mathcal{M}_\theta$ is closely related to the construction of certain interpolation spaces by Calderón [24] and Lozanovskii [89]. Reisner [106] explains Lozanovskii's interpolation results in terminology that is familiar in the West, and also presents some enhancements.

Calderón’s interpolation spaces are Banach lattices of measurable functions on a complete $\sigma$-finite measure space $(\mathcal{X}, \lambda)$. Let $\mathcal{Q}_0$ and $\mathcal{Q}_1$ denote the positive unit balls in Banach function spaces $L_{\rho_0}$ and $L_{\rho_1}$ with function norms $\rho_0$ and $\rho_1$:

$$\mathcal{Q}_i = \{q \mid q \geq 0, \rho_i(q) \leq 1\}, \quad i = 0, 1.$$ 

Calderón [24] introduced the space $L_{\rho_\theta} = L_{\rho_0}^{1-\theta}L_{\rho_1}^\theta$ of measurable functions $q(x)$ such that the inequality $|q(x)| \leq r q_0(x)^{1-\theta}q_1(x)^\theta$ holds for some $r > 0$, some $q_0 \in \mathcal{Q}_0$ and some $q_1 \in \mathcal{Q}_1$. The norm $\rho_\theta(q)$ is then defined as the infimum of all values $r > 0$ that appear in such inequalities. It is obvious that

$$\rho_\theta(q) = \inf \rho_0(q)^{1-\theta}\rho_1(q)^\theta,$$
where the infimum is taken over all nonnegative \( q_0 \in L_{\rho_0} \) and nonnegative \( q_1 \in L_{\rho_1} \) such that \( |q(x)| \leq q_0(x)^{1-\theta} q_1(x)^{\theta} \). If the function norms \( \rho_0 \) and \( \rho_1 \) have the Fatou property (so that \( Q_0 \) and \( Q_1 \) are convex corners), then \( \rho_\theta \) also has the Fatou property. Therefore \( Q_\theta = \{ q \mid q \geq 0, \rho_\theta(q) \leq 1 \} \) is a convex corner as well. The polar of \( Q_\theta \) is equal to the positive unit ball \( M_\theta \) for the associate function seminorm \( \rho_\theta' \):

\[
M_\theta = \{ m \mid m \geq 0, \rho_\theta'(m) \leq 1 \}.
\]

(Here, \( M_\theta \) is a family of densities \( m(x) \) of \( \lambda \)-dominated measures \( m(dx) = m(x) \lambda(dx) \).) The associate function seminorm \( \rho_\theta' \) has the Fatou property, so \( M_\theta \) is a convex corner. It turns out that the associate seminorm \( \rho_\theta' \) can be obtained from \( \rho_\theta \) and \( \rho_1 \) by the same method that was used to construct \( \rho_\theta \) from \( \rho_0 \) and \( \rho_1 \). Thus \( M_\theta \) can be constructed from \( M_0 \) and \( M_1 \) just like \( Q_\theta \) was constructed from \( Q_0 \) and \( Q_1 \).

To illustrate this, let \( Q_0 \) and \( Q_1 \) be the positive unit balls in the Lebesgue spaces \( L_{p_0}(\lambda) \) and \( L_{p_1}(\lambda) \) where \( 1 \leq p_0, p_1 \leq \infty \). Then \( Q_\theta \) is the positive unit ball in \( L_{p_\theta}(\lambda) \), where

\[
1 \over p_\theta = \frac{1 - \theta}{p_0} + \frac{\theta}{p_1}.
\]

The polar \( M_\theta \) of \( Q_\theta \) is the positive unit ball in the conjugate Lebesgue space \( L_{p_\theta'} \) of \( L_{p_\theta} \):

\[
1 \over p_\theta' = \frac{1 - \theta}{p_0} + \frac{\theta}{p_1'} = 1 - \frac{1}{p_\theta'}.
\]

Also interesting is the case when \( Q_0 \) is the positive unit ball of \( L_\infty \), that is the convex corner generated by the constant \( q_0(x) \equiv 1 \). If \( \theta = 1/p \), then \( Q_\theta \) is the convex family of functions \( q_\theta(x) \) such that \( q_\theta^p(x) \) belongs to \( Q_1 \), or such that \( q_\theta(x) = q(x)^\theta \) for some \( q(x) \) in \( Q_1 \). The maximum growth exponent of selections in \( Q_\theta \) vanishes, hence \( W_{Q_\theta}(P) = \theta W_{Q_1}(P) \).

The Banach lattice with function seminorm \( \rho_\theta \) is sometimes called the \( p \)-convexification of the Banach lattice with seminorm \( \rho_1 \). In particular, the Lebesgue space \( L_p(\lambda) \) is the \( p \)-convexification of \( L_1(\lambda) \).

Lozanovskii [90] has formulated a very general interpolation method, and has illustrated its use by constructing all Orlicz spaces as interpolation spaces between \( L^\infty(\lambda) \) and \( L^1(\lambda) \). His method can be described as follows. Let \( A_2^\circ \) denote the class of all positive homogeneous concave functions \( \varphi(q_0, q_1) \) of two arguments \( q_0, q_1 \geq 0 \) such that

\[
\varphi(q_0, 0) = \varphi(0, q_1) = 0, \quad \text{for all } q_0, q_1 \geq 0;
\]

\[
\lim_{q_0 \to \infty} \varphi(q_0, q) = \lim_{q_1 \to \infty} \varphi(q, q_1) = \infty, \quad \text{for all } q > 0.
\]

If \( \varphi(q_0, q_1) \) is a function in \( A_2^\circ \) then so is the conjugate function

\[
\psi(m_0, m_1) = \inf_{q_0, q_1 > 0} \frac{q_0 m_0 + q_1 m_1}{\varphi(q_0, q_1)}, \quad m_0, m_1 \geq 0.
\]

Furthermore, \( \varphi(q_0, q_1) \) is the conjugate of its conjugate \( \psi(m_0, m_1) \), and there is a unique conjugate pair of Young functions \( \Phi(u) \) and \( \Psi(v) \) such that for \( q_1 > 0 \) and \( m_1 > 0 \),

\[
\varphi(q_0, q_1) = q_0 \Phi^{-1}\left( \frac{q_1}{q_0} \right), \quad \psi(m_0, m_1) = m_0 \Psi^{-1}\left( \frac{m_1}{m_0} \right).
\]

50
Now let $Q_0$ and $Q_1$ be convex corners with polars $M_0$ and $M_1$. Given a conjugate pair of functions $\varphi, \psi$ in $A^2$, let $\varphi(Q_0, Q_1)$ denote the class of measurable functions $q(x)$ such that $0 \leq q(x) \leq \varphi(q_0(x), q_1(x))$ for some $q_0(x)$ in $Q_0$ and some $q_1(x)$ in $Q_1$, and similarly define $\psi(M_0, M_1)$. Then $\varphi(Q_0, Q_1)$ and $\psi(M_0, M_1)$ are convex corners and polars of each other.

In particular, if $Q_0$ and $Q_1$ are the positive unit balls in $L^\infty(\lambda)$ and $L^1(\lambda)$, then $\varphi(Q_0, Q_1)$ and $\psi(M_0, M_1)$ are the positive unit balls in the Orlicz spaces $L\Phi(\lambda)$ and $L\Psi(\lambda)$.

In general, let $\varphi_0(q_0, q_1)$ and $\varphi_1(q_0, q_1)$ be functions with conjugates $\psi_0(m_0, m_1)$ and $\psi_1(m_0, m_1)$ in $A^2$, and for $0 < \theta < 1$ let

$$\varphi_\theta(q_0, q_1) = \varphi_0(q_0, q_1)^{1-\theta} \varphi_1(q_0, q_1)^\theta,$$

$$\psi_\theta(m_0, m_1) = \psi_0(m_0, m_1)^{1-\theta} \psi_1(m_0, m_1)^\theta.$$

Then $\varphi_\theta(q_0, q_1)$ and $\psi_\theta(m_0, m_1)$ are conjugate functions in $A^2$. If $Q_0$ and $Q_1$ are convex corners with polars $M_0$ and $M_1$, then the convex corners $\varphi_\theta(Q_0, Q_1)$ and $\psi_\theta(M_0, M_1)$ are obtained by interpolation between the convex corners $\varphi_0(Q_0, Q_1)$ and $\varphi_1(Q_0, Q_1)$, resp. between the convex corners $\psi_0(M_0, M_1)$ and $\psi_1(M_0, M_1)$:

$$\varphi_\theta(Q_0, Q_1) = \varphi_0(Q_0, Q_1)^{1-\theta} \varphi_1(Q_0, Q_1)^\theta,$$

$$\psi_\theta(M_0, M_1) = \psi_0(M_0, M_1)^{1-\theta} \psi_1(M_0, M_1)^\theta.$$

If we set $Q_\varphi = \varphi(Q_0, Q_1)$ and $M_\psi = \psi(M_0, M_1)$, then

$$Q_\varphi = Q_{\varphi_0}^{1-\theta} Q_{\varphi_1}^\theta = Q_{\varphi_0}^{1-\theta} Q_{\varphi_1}^\theta,$$

$$M_\psi = M_{\psi_0}^{1-\theta} M_{\psi_1}^\theta = M_{\psi_0}^{1-\theta} M_{\psi_1}^\theta.$$

In particular, interpolation between the positive unit balls of the Orlicz spaces $L\Phi_0$ and $L\Phi_1$, results in the positive unit ball of the Orlicz space $L\Phi_\theta$, where

$$\Phi_\theta^{-1}(q) = \Phi_0^{-1}(q)^{1-\theta} \Phi_1^{-1}(q)^\theta, \quad q \geq 0.$$

Rao and Ren [?] further discuss this type of interpolation between Orlicz spaces.
VI. THE VALUE OF SIDE INFORMATION

Let \( X \) and \( Y \) be random variables with joint distribution \( P(dx \, dy) = P(dx|y) \, P(dy) \) on the product \( \mathcal{X} \times \mathcal{Y} \) of the measurable spaces \( \mathcal{X} \) and \( \mathcal{Y} \). Let \( \mathcal{Q} \) be a convex family of nonnegative measurable functions on \( \mathcal{X} \). The maximum growth exponent of selections in \( \mathcal{Q} \) when the selections can be made with or without knowledge of the side information \( Y \) may be denoted by \( W(X|Y) \) and \( W(X) \), respectively. The advantage in growth exponent that accrues from knowledge of \( Y \) when making selections in \( \mathcal{Q} \) is equal to the difference

\[
W(X; Y) = W(X|Y) - W(X) \geq 0. \tag{23}
\]

In particular, if \( \mathcal{Q} = \mathcal{Q}(m) \) for some \( \sigma \)-finite measure \( m(dx) \) on \( \mathcal{X} \), then \( W(X; Y) \) is equal to the mutual information

\[
I(X; Y) = I_m(X|Y) - I_m(X) \geq 0.
\]

\( I(X; Y) \) is the advantage in capital growth rate for a gambler who may bet on \( X \) with knowledge of the side information \( Y \) relative to a gambler who must ignore \( Y \). Barron and Cover [6] observed that \( W(X; Y) \leq I(X; Y) \) in the case of log-optimum investment, and that equality holds in the case of log-optimum gambling. We shall put their inequality in a general perspective and find an interpretation for the complement \( I(X; Y) - W(X; Y) \).

One may interpret \( I(X; Y) \) as the total amount of information about \( X \) that can be extracted from the random variable \( Y \). All that information can be exploited by a gambler who bets on \( X \) so as to maximize the growth exponent of compounded wealth, but only a portion \( W(X; Y) \) is useful for maximizing the growth exponent of selections in \( \mathcal{Q} \). It turns out that the complementary portion \( \tilde{W}(X; Y) = I(X; Y) - W(X; Y) \) can be interpreted as the amount of information about \( X \) in \( Y \) that is useful for the dual purpose of minimizing divergence relative to measures in the polar family \( \mathcal{M}(\mathcal{Q}) \).

In Sections A, B, C we consider a \( y \)-convex family \( \Gamma \) of measurable functions \( \gamma(x|y) \geq 0 \) and a \( y \)-convex family \( \Pi \) of transition kernels \( \mu(dx|y) \) such that \( \Pi \) is the \( y \)-conditional polar of \( \Gamma \) and vice versa. We prove two formulas for the mutual information \( I(X; Y) \):

\[
I(X; Y) = W_\Gamma(X; Y) + \tilde{W}_\Gamma(X; Y),
\]

\[
I(X; Y) = \tilde{W}_\Pi(X; Y) + W_\Pi(X; Y).
\]

In general \( \tilde{W}_\Gamma(X; Y) \geq 0 \) and \( \tilde{W}_\Pi(X; Y) \geq 0 \), but there is no guarantee that \( W_\Gamma(X; Y) \geq 0 \) or \( W_\Pi(X; Y) \geq 0 \). The two formulas for \( I(X; Y) \) are generally not the same, but there is a symmetry operation that carries one into the other. In Section D we treat the special case when the two formulas for \( I(X; Y) \) become one and the same. In this special case, \( \Gamma \) and \( \Pi \) are the families of measurable selections \( \gamma(x|y) \) and \( \mu(dx|y) \) in a convex family \( \mathcal{Q} \) and its polar \( \mathcal{M}(\mathcal{Q}) \) given \( y \), and \( I(X; Y) \) is truly decomposed into two complementary terms \( W_\Gamma(X; Y) = \tilde{W}_\Pi(X; Y) \) and \( W_\Pi(X; Y) = \tilde{W}_\Gamma(X; Y) \) that represent the value of the side information \( Y \) for two complementary purposes.

All our results generalize when the random variable \( X \) is defined on a probability space \( (\Omega, \mathcal{F}, P) \) and the side information is represented by a sub-\( \sigma \)-field of \( \mathcal{F} \). The advantage in capital growth rate when betting on \( X \) with knowledge of an information field \( \mathcal{G} \) rather than a sub-\( \sigma \)-field \( \mathcal{H} \) is equal to the conditional mutual information \( I(X; \mathcal{G}|\mathcal{H}) = I_m(X|\mathcal{G}) - \)
\[ I_m(X|\mathcal{H}) \]. Similarly, the advantage in growth exponent when making selections in a convex family \( \mathcal{Q} \) with knowledge of \( \mathcal{G} \) rather than \( \mathcal{H} \) is given by \( W(X; \mathcal{G}|\mathcal{H}) = W(X|\mathcal{G}) - W(X|\mathcal{H}) \).

A. Decomposition of the mutual information \( I(X;Y) \)

Let \( X \) and \( Y \) be random variables with joint distribution \( P(dx\,dy) = P(dx|y)P(dy) \) on the product \( \mathcal{X} \times \mathcal{Y} \) of the measurable spaces \( \mathcal{X} \) and \( \mathcal{Y} \). We consider a \( y \)-convex family \( \Gamma \) of measurable functions \( \gamma(x|y) \geq 0 \) on \( \mathcal{X} \times \mathcal{Y} \). A family is called \( y \)-convex if it is closed under the operation of taking convex combinations with weights that are measurable functions of \( y \). The \( y \)-conditional polar of \( \Gamma \) is defined as the family \( \Pi \) of transition kernels \( \mu(dx|y) \) such that \( \int \gamma(x|y) \mu(dx|y) \leq 1 \) for all \( y \in \mathcal{Y} \) and \( \gamma(x|y) \) in \( \Gamma \). We know that the maximum growth exponent of selections in \( \Gamma \) will not increase if \( \Gamma \) is replaced by the \( y \)-conditional polar of its \( y \)-conditional polar family \( \Pi \). The conditionally log-optimum \( \gamma^*(x|y) \) in \( \Gamma \) and the conditionally log-optimum \( \mu^*(dx|y) \) in \( \Pi \) yield the factorization

\[ P(dx|y) = \gamma^*(x|y) \mu^*(dx|y). \]

Since \( \gamma^*(x|y) = dP/d\mu^*(x|y) \), we see that the maximum conditional growth exponent

\[ W_{\Gamma}(X|Y) = \sup_{\gamma(x|y) \in \Gamma} E\left\{ \log \gamma(X|Y) \right\} = E\left\{ \log \gamma^*(X|Y) \right\} \]

is equal to the minimum conditional information

\[ I_{\Pi}(X|Y) = \inf_{\mu(dx|y) \in \Pi} E \left\{ \log \left( \frac{dP}{d\mu}(X|Y) \right) \right\} = E \left\{ \log \left( \frac{dP}{d\mu^*(X|Y)} \right) \right\}. \]

Let \( \mathcal{Q} = \mathcal{K}(\Gamma) \) denote the convex family of all mixtures \( q^B(x) = \int \gamma(x|y)B(dy) \) where \( B(dy) \) varies over all probability measures on \( \mathcal{Y} \) and \( \gamma(x|y) \) varies over \( \Gamma \). The polar family \( \mathcal{M} = \mathcal{M}(\mathcal{Q}) \) contains all measures \( m(dx) \) such that \( \int \gamma(x|y)m(dx) \leq 1 \) for all \( y \in \mathcal{Y} \) and \( \gamma(x|y) \) in \( \Gamma \). We identify \( \mathcal{M} \) with the set of transition kernels in \( \Pi \) which do not depend on the argument \( y \). The log-optimum \( q^*(x) \) in \( \mathcal{Q} \) and the log-optimum \( m^*(dx) \) in \( \mathcal{M} \) define a factorization of the marginal distribution \( P(dx) \):

\[ P(dx) = q^*(x)m^*(dx). \]

Since \( q^*(x) = dP/dm^*(x) \), the maximum growth exponent

\[ W_{\mathcal{Q}}(X) = \sup_{q(x) \in \mathcal{Q}} E\{ \log q(X) \} = E\{ \log q^*(X) \} \]

is equal to the minimum information

\[ I_{\mathcal{M}}(X) = \inf_{m(dx) \in \mathcal{M}} E \left\{ \log \left( \frac{dP}{dm}(X) \right) \right\} = E \left\{ \log \left( \frac{dP}{dm^*(X)} \right) \right\}. \]

In general, the mixture family \( \mathcal{Q} \) is not a subset of \( \Gamma \) and there is no guarantee that \( W_{\Gamma}(X|Y) \geq W_{\mathcal{Q}}(X) \). Nevertheless, we consider the difference

\[ W_{\Gamma}(X;Y) = W_{\Gamma}(X|Y) - W_{\mathcal{Q}}(X) \]

\[ = E \left\{ \log \left( \frac{\gamma^*(X|Y)}{q^*(X)} \right) \right\} \]

\[ = \sup_{\gamma \in \Gamma} \inf_{q \in \mathcal{Q}} E \left\{ \log \left( \frac{\gamma(X|Y)}{q(X)} \right) \right\}. \]

53
We view $\mathcal{M}$ as a subset of $\Pi$, so $I_{\mathcal{M}}(X|Y) \geq I_{\Pi}(X|Y)$ or $\tilde{W}_\Gamma(X;Y) \geq 0$ where

$$
\tilde{W}_\Gamma(X;Y) = I_{\mathcal{M}}(X|Y) - I_{\Pi}(X|Y)
= E \left\{ \log \left( \frac{\mu^*(dX|Y)}{m^*(dX)} \right) \right\}
= \sup_{\mu \in \Pi} \inf_{m \in \mathcal{M}} E \left\{ \log \left( \frac{\mu(dX|Y)}{m(dX)} \right) \right\}.
$$

Here $\mu(dX|Y)/m(dX)$ denotes the result when the density of $\mu(dx|y)P(dy)$ relative to $m(dx)P(dy)$ is evaluated at $(X,Y)$. If a gambler bets log-optimally on $X$ with knowledge of $Y$ according to the true conditional distribution $P(dx|y)$, then his winnings will grow exponentially with rate $I_{\mathcal{M}}(X|Y)$ or $I_{\Pi}(X|Y)$ depending on whether the odds are set according to the log-optimum $m^*(dx)$ in $\mathcal{M}$ or according to the conditionally log-optimum $\mu^*(dx|y)$ in $\Pi$. Thus $\tilde{W}_\Gamma(X;Y)$ is the advantage in exponential growth rate of wealth for a log-optimum gambler when the bookmaker is not free to set the odds with knowledge of $Y$ according to some conditional measure in $\Pi$, but is forced to ignore $Y$ and post odds according to some measure in the subset $\mathcal{M}$. A rate advantage from the point of view of the gambler is a loss from the point of view of the bookmaker.

**Theorem 20.** Let $X$ and $Y$ be random variables with joint distribution $P(dx\,dy) = P(dx|y)P(dy)$ on $\mathcal{X} \times \mathcal{Y}$. Suppose $\gamma^*(x|y)$ is conditionally log-optimum in a $y$-convex family $\Gamma$ and $\mu^*(dx|y)$ is conditionally log-optimum in the $y$-conditional polar family $\Pi$. If $q^*(x)$ is log-optimum in $Q = K(\Gamma)$ and $m^*(dx)$ is log-optimum in the polar family $M = \mathcal{M}(Q)$, then the following identity holds pointwise with probability one under $P(dx\,dy)$:

$$
\frac{P(dx|y)}{P(dx)} = \frac{\gamma^*(x|y)}{q^*(x)} \frac{\mu^*(dx|y)}{m^*(dx)}.
$$

Taking logarithms and expectations, we obtain the following decomposition of the mutual information $I(X;Y)$:

$$
I(X;Y) = W_\Gamma(X;Y) + \tilde{W}_\Gamma(X;Y).
$$

**Proof:** The theorem follows immediately from the factorizations $P(dx) = q^*(x) m^*(dx)$ and $P(dx|y) = \gamma^*(x|y) \mu^*(dx|y)$. ■

The functionals $W_\Gamma(X;Y)$ and $\tilde{W}_\Gamma(X;Y)$ can be interpreted as the minimum information of $P(dx\,dy)$ relative to kernels in the families $Q\Pi$ and $M\Gamma$, respectively. Indeed, let $G_\Gamma(dx|y)$ and $\tilde{G}_\Gamma(dx|y)$ denote the log-optimum kernels in the families $Q\Pi$ and $M\Gamma$:

$$
G_\Gamma(dx|y) = q^*(x) \mu^*(dx|y) = \frac{q^*(x)}{\gamma^*(x|y)} P(dx|y),
$$

$$
\tilde{G}_\Gamma(dx|y) = m^*(dx) \gamma^*(x|y) = \frac{m^*(dx)}{\mu^*(dx|y)} P(dx|y).
$$

Obviously

$$
\frac{P(dx|y)}{G_\Gamma(dx|y)} = \frac{\gamma^*(x|y)}{q^*(x)}, \quad \frac{P(dx|y)}{\tilde{G}_\Gamma(dx|y)} = \frac{\mu^*(dx|y)}{m^*(dx)}.
$$

54
Taking logarithms and then expectations, we obtain
\[
W_\Gamma(X; Y) = E \left\{ \log \left( \frac{\gamma^*(X|Y)}{q^*(X)} \right) \right\} = I_{G_{\Gamma}}(X|Y) = I_{\mathcal{Q}\Pi}(X|Y),
\]
\[
\tilde{W}_\Gamma(X; Y) = E \left\{ \log \left( \frac{\mu^*(dX|Y)}{m^*(dX)} \right) \right\} = I_{\delta_{\Gamma}}(X|Y) = I_{\mathcal{M}\Gamma}(X|Y).
\]

All transition kernels \( m(dx) \gamma(x|y) \) in \( \mathcal{M}\Gamma \) are subnormalized. Indeed, \( \mathcal{M} \) is regarded as a subset of \( \Pi \) and all transition kernels \( \gamma(x|y) \mu(dx|y) \) in \( \Gamma \Pi \) are subnormalized because the \( y \)-conditional polar condition \( \int \gamma(x|y) \mu(dx|y) \leq 1 \) holds for all \( \gamma(x|y) \) in \( \Gamma \) and \( \mu(dx|y) \) in \( \Pi \). The normalized transition kernels in \( \mathcal{M}\Gamma \) are exactly those for which the side information \( Y \) is irrelevant when minimizing the information relative to transition kernels in \( \Pi \). Indeed, suppose \( \tilde{G}(dx|y) \) is a conditional distribution under which the conditionally log-optimum selection in \( \Pi \) given \( y \) is independent of \( y \) and therefore equal to some measure \( m(dx) \) in \( \mathcal{M} \). Analogous to the factorization \( P(dx|y) = \gamma^*(x|y) \mu^*(dx|y) \), we see that \( \tilde{G}(dx|y) = m(dx) \gamma(x|y) \) where \( \gamma(x|y) \) is conditionally log-optimum given \( y \) in \( \Gamma \) under \( G(dx|y) \). Thus \( \tilde{G}(dx|y) \) is a normalized transition kernel in \( \mathcal{M}\Gamma \). Conversely, if \( \tilde{G}(dx|y) = m(dx) \gamma(x|y) \) is a normalized transition kernel in \( \mathcal{M}\Gamma \) then \( m(dx) \) is conditionally log-optimum in \( \Pi \) given \( y \) under \( \tilde{G}(dx|y) \) because the conditional KKT conditions hold for all \( \mu(dx|y) \) in \( \Pi \):
\[
\int \frac{\mu(dx|y)}{m(dx)} \tilde{G}(dx|y) = \int \frac{\mu(dx|y)}{m(dx)} m(dx) \gamma(x|y) \leq \int \mu(dx|y) \gamma(x|y) \leq 1.
\]

In general \( \tilde{W}_\Gamma(X; Y) \geq 0 \) or \( W_\Gamma(X; Y) \leq I(X; Y) \), but it is not always true that \( W_\Gamma(X; Y) \geq 0 \) or \( \tilde{W}_\Gamma(X; Y) \leq I(X; Y) \). The graphical interpretation of \( \tilde{W}_\Gamma(X; Y) \) which will be mentioned at the end of Section VII.C below will confirm this. The inequality \( W_\Gamma(X; Y) \leq I(X; Y) \) can be proved directly by observing that \( P(dx|y) = q^*(x) m^*(dx) \) is log-optimum in \( \mathcal{Q}\mathcal{M} \) and \( \mathcal{Q}\mathcal{M} \subset \mathcal{Q}\Pi \) (because \( \mathcal{M} \) is viewed as a subset of \( \Pi \)), hence
\[
W_\Gamma(X; Y) = I_{\mathcal{Q}\Pi}(X|Y) \leq I_{\mathcal{Q}\mathcal{M}}(X|Y) = I(X; Y).
\]

We have equality \( W_\Gamma(X; Y) = I(X; Y) \) or \( \tilde{W}_\Gamma(X; Y) = 0 \) iff the conditionally log-optimum \( \mu^*(dx|y) \) in \( \Pi \) given \( y \) is independent of \( y \) and therefore equal to the log-optimum \( m^*(dx) \) in \( \mathcal{M} \). The conditionally log-optimum \( G_\Gamma(dx|y) = q^*(x) \mu^*(dx|y) \) in \( \mathcal{Q}\Pi \) is then equal to \( q^*(x) m^*(dx) = P(dx) \), and the conditionally log-optimum \( \tilde{G}_\Gamma(dx|y) = m^*(dx) \gamma^*(x|y) \) in \( \mathcal{M}\Gamma \) is equal to \( \mu^*(dx|y) \gamma^*(x|y) = P(dx|y) \). Thus the equalities \( W_\Gamma(X; Y) = I(X; Y) \) and \( \tilde{W}_\Gamma(X; Y) = 0 \) are really manifestations of the pointwise equalities
\[
\frac{P(dx|y)}{G_\Gamma(dx|y)} = \frac{P(dx|y)}{P(dx)}, \quad \frac{P(dx|y)}{\tilde{G}_\Gamma(dx|y)} = 1.
\]

**B. Optimization of test channels**

Again we consider a \( y \)-convex family \( \Gamma \) and its \( y \)-conditional polar family \( \Pi \), the mixture family \( \mathcal{Q} = \mathcal{K}(\Gamma) \) and its polar \( \mathcal{M} = \mathcal{M}(\mathcal{Q}) \). Let \( X \) be a random variable with marginal distribution \( P(dx) \) such that \( W_\mathcal{Q}(X) \) is well defined and finite. For any joint distribution \( P(dx dy) = P(dx) F(dy|x) \) on \( X \times Y \) we have the inequalities
\[
W_\Gamma(X|Y) - W_\mathcal{Q}(X) = W_\Gamma(X; Y) \leq I(X; Y), \tag{24}
\]
0 \leq \tilde{W}_\Gamma(X; Y) = I_\mathcal{M}(X) - I_\Pi(X|Y) + I(X; Y).

We argue that for a suitable test channel \( F_\Gamma(dy|x) \), we have equality \( \tilde{W}_\Gamma(X; Y) = 0 \) and \( W_\Gamma(X; Y) = I(X; Y) \) under \( P(dx) F_\Gamma(dy|x) \).

**Theorem 21.** Let \( \Gamma, \Pi, \mathcal{Q} = \mathcal{K}(\Gamma), \mathcal{M} = \mathcal{M}(\mathcal{Q}) \) be specified as above. If \( X \) is a random variable with marginal distribution \( P(dx) \) such that \( W_\mathcal{Q}(X) \) is well defined and finite, then

\[
\sup_{F(dy|x)} [W_\Gamma(X|Y) - I(X; Y)] = W_\mathcal{Q}(X)
\]

or equivalently, since \( W_\mathcal{Q}(X) = I_\mathcal{M}(X) \) and \( W_\Gamma(X|Y) = I_\Pi(X|Y) \),

\[
\sup_{F(dy|x)} [I_\Pi(X|Y) - I(X; Y)] = I_\mathcal{M}(X).
\]

A test channel \( F(dy|x) \) attains the suprema on the left hand sides iff \( W_\Gamma(X; Y) = I(X; Y) \) or equivalently iff \( \tilde{W}_\Gamma(X; Y) = 0 \) under \( P(dx) F(dy|x) \). Such a test channel can be obtained by taking a function \( \gamma_\Gamma(x|y) \) in \( \Gamma \) and a probability measure \( B_\Gamma(dy) \) on \( \mathcal{Y} \) such that the mixture \( q^*(x) = \int \gamma_\Gamma(x|y) B_\Gamma(dy) \) is log-optimum in \( \mathcal{Q} \), and setting

\[
F_\Gamma(dy|x) = \frac{\gamma_\Gamma(x|y)}{q^*(x)} B_\Gamma(dy).
\]

**Proof:** The inequality (24) implies that \( \sup_{F(dy|x)} [W_\Gamma(X|Y) - I(X; Y)] \leq W_\mathcal{Q}(X) \). To prove the theorem it suffices to argue that (24) holds with equality under \( P(dx) F_\Gamma(dy|x) \).

By construction, \( F_\Gamma(dy|x) \) is a normalized transition kernel from \( \mathcal{X} \) to \( \mathcal{Y} \). Thus \( F_\Gamma(dy|x) \) combined with \( P(dx) \) defines a joint probability distribution

\[
P_\Gamma(dx \, dy) = P(dx) F_\Gamma(dy|x) = \frac{P(dx) \gamma_\Gamma(x|y) B_\Gamma(dy)}{q^*(x)}.
\]

Now \( P(dx)/q^*(x) \) is exactly the log-optimum measure \( m^*(dx) \) in \( \mathcal{M} \), so

\[
P_\Gamma(dx \, dy) = m^*(dx) \gamma_\Gamma(x|y) B_\Gamma(dy).
\]

The marginal distribution of \( Y \) under \( P_\Gamma(dx \, dy) \) is equal to \( B_\Gamma(dy) \) and the conditional distribution of \( X \) given \( Y \) is defined by the subnormalized transition kernel

\[
P_\Gamma(dx|y) = m^*(dx) \gamma_\Gamma(x|y).
\]

Now \( P_\Gamma(dx|y) \) has a unique factorization \( \mu^*(dx|y) \gamma^*(x|y) \) in \( \Pi \Gamma \), so \( m^*(dx) \) is equal to the conditionally log-optimum \( \mu^*(dx|y) \) in \( \Pi \) and \( \gamma_\Gamma(x|y) \) is equal to the conditionally log-optimum \( \gamma^*(x|y) \) in \( \Gamma \). The conditionally log-optimum \( G_\Pi(dx|y) = m^*(dx) \gamma^*(x|y) \) in \( \mathcal{M} \Gamma \) is equal to the conditional distribution \( P_\Gamma(dx|y) = m^*(dx) \gamma_\Gamma(x|y) \) and the conditionally log-optimum \( \tilde{G}_\Gamma(dx|y) = q^*(x) \mu^*(dx|y) \) in \( \mathcal{Q} \Pi \) is equal to \( P(dx) = q^*(x) m^*(dx) \). Thus (24) holds with equality under \( P(dx) F_\Gamma(dy|x) \), and the theorem follows.

Suppose \( X \) and \( Y \) have joint distribution \( P(dx \, dy) = P(dx) F(dy|x) \). Since \( I(X; Y) = W_\Gamma(X; Y) + \tilde{W}_\Gamma(X; Y) \) and \( W_\Gamma(X; Y) = W_\Gamma(X|Y) - W_\mathcal{Q}(X) \), we obtain the identity

\[
W_\mathcal{Q}(X) = [W_\Gamma(X|Y) - I(X; Y)] + \tilde{W}_\Gamma(X; Y).
\]
The left hand side $W_\Omega(X)$ is not a function of $F(dy|x)$, hence a test channel $F(dy|x)$ will maximize the term $[W_T(X|Y) - I(X;Y)]$ if it minimizes the complementary term $\bar{W}_T(X;Y)$. By Theorem 21, the maximum of $[W_T(X|Y) - I(X;Y)]$ over all test channels $F(dy|x)$ is equal to $W_\Omega(X)$ and the minimum of $\bar{W}_T(X;Y)$ over all test channels $F(dx|y)$ vanishes. Finding a forward test channel that attains the maximum $W_\Omega(X) = \sup_{F(dy|x)} [W_T(X|Y) - I(X;Y)]$ or that attains the infimum $0 = \inf_{F(dy|x)} \bar{W}_T(X;Y)$ amounts to finding a joint probability distribution $P_T(dx,dy) = P(dx) F_T(dy|x) = F_T(dy|x) B_T(dy)$ such that

$$P_T(dx,dy) = P(dx) F_T(dy|x) = \bar{G}_T(dx|y) B_T(dy)$$

for some backward test channel $\bar{G}_T(dx|y) = m^*(dx) \gamma_T(x|y)$ in $\mathcal{M} \Gamma$. The output distribution $\int_x P(dx) F_T(dy|x)$ of the forward test channel $F_T(dy|x)$ is equal to the marginal distribution $B_T(dy)$ and the output distribution $\int_x \bar{G}_T(dx|y) B_T(dy)$ of the backward test channel $\bar{G}_T(dx|y)$ must be equal to the marginal $P(dx)$.

C: Another decomposition defined in a parallel fashion

We develop a theory parallel to that in Sections A, B. Our starting point is a $y$-convex family $\Pi$ of transition kernels from $\mathcal{Y}$ to $\mathcal{X}$ and the $y$-conditional polar family $\Gamma$ of non-negative measurable functions $\gamma(x|y)$ such that $\int \gamma(x|y) \mu(dx|y) \leq 1$ for all $\mu(dx|y)$ in $\Pi$. We now consider the family $\mathcal{M} = \mathcal{K}(\Pi)$ of all mixtures $m_B^\mathcal{M}(dx) = \int \mu(dx|y) B(dy)$ where $\mu(dx|y)$ ranges over $\Pi$ and $B(dy)$ ranges over all probability measures on $\mathcal{Y}$, and the polar family $\mathcal{Q} = \mathcal{Q}(\mathcal{M})$ of functions $q(x) \geq 0$ such that $\int q(x) \mu(dx|y) \leq 1$ for all $\mu(dx|y)$ in $\Pi$. We identify $\mathcal{Q}$ with the set of functions in $\Gamma$ which do not depend on the argument $y$. The log-optimum $q^*(x)$ in $\mathcal{Q}$ and $m^*(dx)$ in $\mathcal{M}$ again define a factorization $P(dx) = q^*(x) m^*(dx)$, but the reader should realize that this is not the same factorization as in Section A because we are now dealing with the mixture family $\mathcal{M} = \mathcal{K}(\Pi)$ and its polar $\mathcal{Q} = \mathcal{Q}(\mathcal{M})$ rather than the mixture family $\mathcal{Q} = \mathcal{K}(\Gamma)$ and its polar $\mathcal{M} = \mathcal{M}(\mathcal{Q})$. The special case where the new definitions of $\mathcal{Q}$ and $\mathcal{M}$ agree with the earlier ones is very interesting and will be discussed in the next section.

Given this new setup, we define $W_\Pi(X;Y)$, $\bar{W}_\Pi(X;Y)$, $G_\Pi(dx|y)$ and $\bar{G}_\Pi(dx|y)$ like we defined $\bar{W}_T(X;Y)$, $W_T(X;Y)$, $\bar{G}_T(dx|y)$ and $G_T(dx|y)$ previously. Thus we set

$$W_\Pi(X;Y) = I_{\mathcal{M}}(X|Y) - I_{\Pi}(X|Y)$$

$$= E \left\{ \log \left( \frac{\mu^*(dX|Y)}{m^*(dX)} \right) \right\}$$

$$= \sup_{\mu \in \Pi} \inf_{m \in \mathcal{M}} E \left\{ \log \left( \frac{\mu(dx|Y)}{m(dx)} \right) \right\},$$

$$\bar{W}_\Pi(X;Y) = W_T(X|Y) - W_\Omega(X)$$

$$= E \left\{ \log \left( \frac{\gamma^*(X|Y)}{q^*(X)} \right) \right\}$$

$$= \sup_{\gamma \in \Gamma} \inf_{q \in \mathcal{Q}} E \left\{ \log \left( \frac{\gamma(X|Y)}{q(X)} \right) \right\}.$$
made in \( \Gamma \) with rather than without knowledge of the side information \( Y \). In general \( \mathcal{M} \) is not a subset of \( \Pi \), so it is not always true that \( I_{\mathcal{M}}(X|Y) \geq I_{\Pi}(X|Y) \) or \( W_{\Pi}(X;Y) \geq 0 \).

Since \( P(dx|y) = \gamma^*(x|y) \mu^*(dx|y) \) and \( P(dx) = q^*(x) m^*(dx) \), we have the identity
\[
\frac{P(dx|y)}{P(dx)} = \frac{\gamma^*(x|y) \mu^*(dx|y)}{q^*(x) m^*(dx)}
\]

and the following decomposition of the mutual information \( I(X;Y) \):
\[
I(X;Y) = \tilde{W}_{\Pi}(X;Y) + W_{\Pi}(X;Y).
\]

The conditionally log-optimum selections in \( \mathcal{M} \Gamma \) and \( \mathcal{Q} \Pi \) are now denoted by
\[
G_{\Pi}(dx|y) = m^*(dx) \gamma^*(x|y) = \frac{m^*(dx)}{\mu^*(dx|y)} P(dx|y),
\]
\[
\tilde{G}_{\Pi}(dx|y) = q^*(x) \mu^*(dx|y) = \frac{q^*(x)}{\gamma^*(x|y)} P(dx|y).
\]
The functionals \( W_{\Pi}(X;Y) \) and \( \tilde{W}_{\Pi}(X;Y) \) can be viewed as the minimum information relative to selections in \( \mathcal{M} \Gamma \) and \( \mathcal{Q} \Pi \), respectively:
\[
W_{\Pi}(X;Y) = E \left\{ \log \left( \frac{\mu^*(X|Y)}{m^*(X)} \right) \right\} = I_{G_{\Pi}}(X|Y) = I_{\mathcal{M} \Gamma}(X|Y),
\]
\[
\tilde{W}_{\Pi}(X;Y) = E \left\{ \log \left( \frac{\gamma^*(X|Y)}{q^*(X)} \right) \right\} = I_{\tilde{G}_{\Pi}}(X|Y) = I_{\mathcal{Q} \Pi}(X|Y).
\]

All transition kernels \( q(x) \mu(dx|y) \) in \( \mathcal{Q} \Pi \) are subnormalized, so \( \tilde{W}_{\Pi}(X;Y) \geq 0 \) or \( W_{\Pi}(X;Y) \leq I(X;Y) \). The normalized kernels in \( \mathcal{Q} \Pi \) are the conditional distributions under which the conditionally log-optimum selection in \( \Gamma \) given \( y \) does not depend on \( y \) and therefore is a function \( q(x) \) in \( \mathcal{Q} \). The transition kernels \( m(dx) \gamma(x|y) \) in \( \mathcal{M} \Gamma \) need not be subnormalized, so there is no guarantee that \( W_{\Pi}(X;Y) \geq 0 \) or \( \tilde{W}_{\Pi}(X;Y) \leq I(X;Y) \).

The measure \( P(dx) = q^*(x) m^*(dx) \) is log-optimum in \( \mathcal{Q} \mathcal{M} \). Now \( \mathcal{Q} \) is regarded as a subset of \( \Gamma \), so \( \mathcal{Q} \mathcal{M} \subseteq \Gamma \mathcal{M} \) and
\[
W_{\Pi}(X;Y) = I_{\Gamma \mathcal{M}}(X|Y) \leq I_{\mathcal{Q} \mathcal{M}}(X|Y) = I(X;Y).
\]
We have equality \( W_{\Pi}(X;Y) = I(X;Y) \) or \( \tilde{W}_{\Pi}(X;Y) = 0 \) or \( W_{\Gamma}(X|Y) = W_{\mathcal{Q}}(X) \) if the conditionally log-optimum \( \gamma^*(x|y) \) in \( \Gamma \) given \( y \) is independent of \( y \) and therefore equal to the log-optimum \( q^*(x) \) in \( \mathcal{Q} \). In that case, \( \tilde{G}_{\Pi}(dx|y) = q^*(x) \mu^*(x|y) \) is equal to the conditional distribution \( P(dx|y) = \gamma^*(x|y) \mu^*(dx|y) \), and \( G_{\Pi}(dx|y) = m^*(dx) \gamma^*(x|y) \) is equal to \( P(dx) = m^*(dx) q^*(x) \). Therefore we have the pointwise identities
\[
\frac{P(dx|y)}{G_{\Pi}(dx|y)} = \frac{P(dx|y)}{P(dx)}, \quad \frac{P(dx|y)}{\tilde{G}_{\Pi}(dx|y)} = 1.
\]

We now draw a parallel with Theorem 21. For any marginal distribution \( P(dx) \) such that \( I_{\mathcal{M}}(X) \) is well defined and finite we construct a test channel \( F_{\Pi}(dy|x) \) such that the inequalities \( W_{\Pi}(X;Y) \geq 0 \) and \( W_{\Pi}(X;Y) \leq I(X;Y) \) hold with equality under \( P(dx) F_{\Pi}(dy|x) \).

58
Theorem 22. Let $\Pi, \Gamma, \mathcal{M} = \mathcal{K}(\Pi), \mathcal{Q} = \mathcal{Q}(\mathcal{M})$ be as specified. If $X$ is a random variable with marginal distribution $P(dx)$ such that $I(\mathcal{M})$ is well defined and finite, then

$$I(\mathcal{M}) = \inf_{F(dy|x)} I_n(X|Y),$$

or equivalently, since $I(\mathcal{M}) = W(\mathcal{Q})$ and $I_n(X|Y) = W_\Gamma(X|Y)$,

$$W(\mathcal{Q}) = \inf_{F(dy|x)} W_\Gamma(X|Y).$$

A test channel $F(dy|x)$ attains these infima iff $\tilde{W}(\Pi ; X, Y) = 0$ or equivalently iff $W_\Gamma(X; Y) = I(X; Y)$ under $P(dx) F(dy|x)$. Such a test channel can be obtained by taking a transition kernel $\mu_\Pi(dx|y)$ in $\Pi$ and a probability measure $B_\Pi(dy)$ on $\mathcal{Y}$ such that the mixture $m^*(dx) = \int \mu_\Pi(dx|y) B_\Pi(dy)$ is log-optimum in $\mathcal{M}$, and setting

$$F_\Pi(dy|x) = \frac{\mu_\Pi(dx|y)}{m^*(dx)} B_\Pi(dy).$$

Since $q^*(x) = P(dx)/m^*(dx)$, the joint distribution $P(dx) F_\Pi(dy|x)$ has the form

$$P_\Pi(dx, dy) = P(dx) F_\Pi(dy|x) = q^*(x) \mu_\Pi(dx|y) B_\Pi(dy).$$

If $X$ and $Y$ have this joint distribution then the marginal distribution of $Y$ is equal to $B_\Pi(dy)$ and the conditional distribution of $X$ given $Y$ is defined by the transition kernel

$$P_\Pi(dx|y) = q^*(x) \mu_\Pi(dx|y).$$

There is a unique factorization $P_\Pi(dx|y) = \gamma^*(x|y) \mu^*(dx|y)$ in $\Gamma \Pi$, so $q^*(x)$ is equal to the conditionally log-optimum $\gamma^*(x|y)$ in $\Gamma$ and $\mu_\Pi(dx|y)$ is equal to the conditionally log-optimum $\mu^*(dx|y)$ in $\Pi$. Thus the conditionally log-optimum $\tilde{G}_\Pi(dx|y) = q^*(x) \mu^*(dx|y)$ in $\mathcal{Q} \Pi$ is equal to the conditional distribution $P_\Pi(dx|y) = q^*(x) \mu_\Pi(dx|y)$ and the conditionally log-optimum $G_\Pi(dx|y) = m^*(dx) \gamma^*(x|y)$ in $\mathcal{M} \Gamma$ is equal to $m^*(dx) q^*(x) = P(dx)$. The output distribution $\int_{\mathcal{X}} P(dx) F_\Pi(dy|x)$ of the forward channel $F_\Pi(dy|x)$ is equal to the marginal $B_\Pi(dy)$ and the output distribution $\int_{\mathcal{Y}} \tilde{G}_\Pi(dx|y) B_\Pi(dy)$ of the backward channel $\tilde{G}_\Pi(dx|y)$ is equal to the marginal $P(dx)$.

There is an obvious parallelism between this section and Sections A, B. In fact, these parallel treatments are completely equivalent to each other. Indeed, consider a $y$-convex family $\Pi$, its $y$-conditional polar $\Gamma$, the mixture family $\mathcal{M} = \mathcal{K}(\Pi)$ and its polar $\mathcal{Q} = \mathcal{Q}(\mathcal{M})$. We can define a $y$-convex family $\tilde{\Gamma}$, its $y$-conditional polar $\tilde{\Pi}$, the mixture family $\tilde{\mathcal{Q}} = \mathcal{K}(\tilde{\Gamma})$ and its polar $\tilde{\mathcal{M}} = \mathcal{M}(\tilde{\mathcal{Q}})$ such that

$$W_\Pi(X; Y) = E \left\{ \log \left( \frac{\mu^*(dX|Y)}{m^*(dX)} \right) \right\} = E \left\{ \log \left( \frac{\gamma^*(X|Y)}{\tilde{q}^*(X)} \right) \right\} = W_\Gamma(X; Y),$$

$$\tilde{W}_\Pi(X; Y) = E \left\{ \log \left( \frac{\gamma^*(X|Y)}{\tilde{q}^*(X)} \right) \right\} = E \left\{ \log \left( \frac{\tilde{\mu}^*(dX|Y)}{\tilde{m}^*(dX)} \right) \right\} = \tilde{W}_\Gamma(X; Y).$$

To see this, pick a $\sigma$-finite measure $\lambda(dx)$ such that $I_\lambda(X)$ is well defined and finite (the natural choice is $\lambda(dx) = P(dx)$). Let $\tilde{\Gamma}$ consist of all densities $\tilde{\gamma}(x|y)$ of the form $\mu(dx|y)/\lambda(dx)$.
for some $\mu(dx|y)$ in $\Pi$, and let $\bar{\Pi}$ consist of all kernels $\bar{\mu}(dx|y)$ of the form $\gamma(x|y)\lambda(dx)$ for some $\gamma(x|y)$ in $\Gamma$. Also, let $\bar{\mathcal{M}}$ consist of all measures $\bar{m}(dx)$ of the form $q(x)\lambda(dx)$ for some $q(x)$ in $\mathcal{Q}$, and let $\bar{\mathcal{Q}}$ consist of all densities $\bar{q}(x)$ of the form $m(dx)/\lambda(dx)$ for some $m(dx)$ in $\mathcal{M}$. By construction $\mathcal{M}\Gamma = \bar{\mathcal{Q}}\bar{\Pi}$ and $\mathcal{Q}\Pi = \bar{\mathcal{M}}\bar{\Gamma}$. The conditionally log-optimum selections in these families admit the representations

$$G_\Pi(dx|y) = m^*(dx)\gamma^*(x|y) = \bar{q}^*(x)\bar{\mu}^*(dx|y) = \bar{G}_\Gamma(dx|y),$$

$$\bar{G}_\Pi(dx|y) = q^*(x)\mu^*(x|y) = \bar{m}^*(dx)\bar{\gamma}^*(x|y) = \bar{G}_\Gamma(dx|y).$$

A test channel $F_\Pi(dy|x)$ such that $W_\Pi(X;Y) = 0$ or $W_\Pi(X;Y) = I(X;Y)$ under $P(dx)F_\Pi(dx|y)$ also functions as a test channel $F_\Gamma(dy|x)$ such that $W_\Gamma(X;Y) = 0$ or $W_\Gamma(X;Y) = I(X;Y)$ under $P(dx)F_\Gamma(dy|x)$. In fact, $B_\Pi(dy) = B_\Gamma(dy)$ and $\mu_\Pi(dx|y) = \gamma_\Gamma(x|y)\lambda(dx)$, hence

$$F_\Pi(dy|x) = \frac{\mu_\Pi(dx|y)}{m^*(dx)}B_\Pi(dy) = \frac{\gamma_\Gamma(x|y)}{q^*(x)}B_\Gamma(dy) = F_\Gamma(dy|x),$$

$$P_\Pi(dx\,dy) = q^*(x)\mu_\Pi(dx|y)B_\Pi(dy) = \bar{m}^*(dx)\bar{\gamma}_\Gamma(x|y)B_\Gamma(dy) = P_\Gamma(dx\,dy).$$

The correspondence between $\Pi, \Gamma, \mathcal{M}, \mathcal{Q}$ and $\bar{\Pi}, \bar{\Gamma}, \bar{\mathcal{Q}}, \bar{\mathcal{M}}$ is involutive, so the problems in Sections A, B and the parallel problems in this section are transformed back and forth into each other.

## D. A special case with perfect symmetry

Let $\mathcal{Q}$ and $\mathcal{M}$ be convex families of nonnegative measurable functions $q(x)$ and measures $m(dx)$ such that $\mathcal{M}$ is the polar of $\mathcal{Q}$ and vice versa. Let $\Gamma$ denote the $y$-convex family of all measurable selections $\gamma(x|y)$ in $\mathcal{Q}$ given $y$ and let $\Pi$ denote the $y$-convex family of all measurable selections $\mu(dx|y)$ in $\mathcal{M}$ given $y$. Then $\Gamma$ and $\Pi$ are $y$-conditional polars of each other and

$$\mathcal{Q} = \mathcal{K}(\Gamma), \quad \mathcal{M} = \mathcal{K}(\Pi).$$

The decompositions of $I(X;Y)$ in Sections A and C are different in general, but they are identical when the mixture families $\mathcal{Q} = \mathcal{K}(\Gamma)$ and $\mathcal{M} = \mathcal{K}(\Pi)$ are polars of each other. In this special case we have the identities

$$I(X;Y) = W_\Gamma(X;Y) + W_\Pi(X;Y),$$

$$W_\Gamma(X;Y) = E\left\{ \log \left( \frac{\gamma^*(X|Y)}{q^*(X)} \right) \right\} = W_\Pi(X;Y) \geq 0,$$

$$W_\Pi(X;Y) = E\left\{ \log \left( \frac{\mu^*(dX|Y)}{m^*(dX)} \right) \right\} = W_\Gamma(X;Y) \geq 0.$$

Both $W_\Gamma(X;Y) \geq 0$ and $W_\Pi(X;Y) \geq 0$ since $\mathcal{Q}$ and $\mathcal{M}$ can be identified with subsets of $\Gamma$ and $\Pi$ (namely, the measurable selections given $y$ which do not depend on $y$).
The complementary terms in the decomposition \( I(X;Y) = W_{\Gamma}(X;Y) + W_{\Pi}(X;Y) \) express the amount of information about the random variable \( X \) that can be extracted from the side information \( Y \) for two complementary purposes. The term \( W_{\Gamma}(X;Y) \) represents the increment in growth exponent when selections are made in \( \mathcal{Q} \) with rather than without knowledge of \( Y \). Similarly, \( W_{\Pi}(X;Y) \) expresses the decrement in information relative to selections that are made in \( \mathcal{M} \) with rather than without knowledge of \( Y \).

All kernels \( q(x) \mu(dx|y) \) in \( \mathcal{Q} \) and all kernels \( m(dx) \gamma(x|y) \) in \( \mathcal{M} \) are subnormalized. If the log-optimum \( G_{\Gamma}(dx|y) = q^*(x) \mu^*(dx|y) = G_{\Pi}(dx|y) \) in \( \mathcal{Q} \) is normalized then \( W_{\Gamma}(X;Y) = I_{\mathcal{Q}\Pi}(X|Y) = W_{\Pi}(X;Y) \) is the minimum divergence of \( P(dx|dy) \) relative to joint distributions \( q(x) \mu(dx|y) B(dy) \) under which the conditionally log-optimum selection in \( \mathcal{Q} \) given \( y \) does not depend on \( y \). Similarly, if the log-optimum \( G_{\Pi}(dx|y) = m^*(dx) \gamma^*(x|y) = G_{\Gamma}(dx|y) \) in \( \mathcal{M} \) is normalized then \( W_{\Pi}(X;Y) = I_{\mathcal{M}\Pi}(X|Y) = W_{\Gamma}(X;Y) \) is the minimum divergence relative to joint distributions \( m(dx) \gamma(x|y) B(dy) \) under which the conditionally log-optimum selection in \( \mathcal{M} \) given \( y \) does not depend on \( y \). Clearly \( W_{\Gamma}(X;Y) = I(X;Y) \) and \( W_{\Pi}(X;Y) = 0 \) under \( P_{\Gamma}(dx|dy) = P(dx) F_{\Gamma}(dy|x) \), whereas \( W_{\Gamma}(X;Y) = 0 \) and \( W_{\Pi}(X;Y) = I(X;Y) \) under \( P_{\Pi}(dx|dy) = P(dx) F_{\Pi}(dy|x) \).

The balance between the two complementary terms in the decomposition \( I(X;Y) = W_{\Gamma}(X;Y) + W_{\Pi}(X;Y) \) can be tilted to any intermediate position between the two extremes 0 and \( I(X;Y) \). Indeed, let \( \lambda(dx) \) be a \( \sigma \)-finite reference measure on \( \mathcal{X} \) such that \( P(dx) \) has density \( p(x) \) relative to \( \lambda(dx) \) and \( P(dx|y) \) has density \( p(x|y) \) relative to \( \lambda(dx) \) \( P(dy) \). For \( 0 < \theta < 1 \) define \( \mathcal{Q}^{\theta} \) as the family of densities \( q^{(\theta)}(x) \) of measures \( Q^{(\theta)}(dx)^{\theta} \lambda(dx)^{1-\theta} = q^{(\theta)}(x) \lambda(dx) \) such that \( Q(dx) = q(x) \lambda(dx) \) is subnormalized, and define \( \mathcal{M}^{\theta} \) as the family of measures \( M(dx)^{1-\theta} \lambda(dx)^{\theta} = m(x)^{1-\theta} \lambda(dx) \) such that \( M(dx) = m(x) \lambda(dx) \) is subnormalized. In particular, \( \mathcal{Q}^{0} \) is the convex corner that is generated by the constant function \( q^{(0)}(x) \equiv 1 \) and \( \mathcal{M}^{0} \) contains all \( \lambda \)-dominated sub-probability measures \( dM = m \lambda(dx) \), whereas \( \mathcal{Q}^{1} = \mathcal{Q}(\lambda) \) contains all densities of \( \lambda \)-dominated subprobability measures \( dQ = q \lambda(dx) \) and \( \mathcal{M}^{1} \) is the convex corner that is generated by \( \lambda \). One may view \( \mathcal{Q}^{\theta} \) as the family of densities of measures in \( \mathcal{M}^{1-\theta} \) and \( \mathcal{M}^{\theta} \) as the family of measures with density in \( \mathcal{Q}^{1-\theta} \). Observe that \( r = \theta^{-1} \) and \( s = (1-\theta)^{-1} \) are conjugate exponents such that \( r^{-1} + s^{-1} = 1 \). The spaces \( \mathcal{Q}^{\theta} \) and \( \mathcal{M}^{\theta} \) result from intersecting the positive cone with the closed unit ball in the spaces \( L^{r}(\lambda) \) and \( L^{s}(\lambda) \), and Hölder's inequality implies that \( \mathcal{Q}^{\theta} \) and \( \mathcal{M}^{\theta} \) are convex corners and polars of each other. Clearly, \( p(x) \) is log-optimum in \( \mathcal{Q}^{0} \) and \( p(x)^{1-\theta} \lambda(dx) \) is log-optimum in \( \mathcal{M}^{\theta} \). Similarly, \( p(x|y) \) and \( p(x|y)^{1-\theta} \lambda(dx) \) are conditionally log-optimum given \( y \) in the families \( \Gamma^{\theta} \) and \( \Pi^{\theta} \) of conditional selections given \( y \) in \( \mathcal{Q}^{\theta} \) and \( \mathcal{M}^{\theta} \). It follows that

\[
W_{\Gamma^{\theta}}(X;Y) = \theta I(X;Y), \quad W_{\Pi^{\theta}}(X;Y) = (1-\theta) I(X;Y).
\]

Generally speaking, if \( \mathcal{Q}^{\theta} \) interpolates between convex families \( \mathcal{Q}^{0} \) and \( \mathcal{Q}^{1} \), then \( \mathcal{Q}^{\theta} \) is convex and its polar \( \mathcal{M}^{\theta} \) is the family \( \mathcal{M}^{\theta} \) that interpolates between the polar families \( \mathcal{M}^{0} = \mathcal{M}(\mathcal{Q}^{0}) \) and \( \mathcal{M}^{1} = \mathcal{M}(\mathcal{Q}^{1}) \). Also, the family \( \Gamma^{\theta} \) that interpolates between the families \( \Gamma^{0} \) and \( \Gamma^{1} \) of \( y \)-conditional selections in \( \mathcal{Q}^{0} \) and \( \mathcal{Q}^{1} \) is the \( y \)-convex family of all \( y \)-conditional selections in \( \mathcal{Q}^{\theta} \), and the family \( \Pi^{\theta} \) that interpolates between the \( y \)-conditional polar families \( \Pi^{0} \) and \( \Pi^{1} \) of \( \Gamma^{0} \) and \( \Gamma^{1} \) is equal to both the family of \( y \)-conditional selections in \( \mathcal{M}^{\theta} \) given \( y \) and the \( y \)-conditional polar family of \( \Gamma^{\theta} \). The functionals \( W_{\mathcal{Q}^{\theta}}(X), \ W_{\Gamma^{\theta}}(X|Y), \ I_{\mathcal{M}^{\theta}}(X) \) and \( I_{\Pi^{\theta}}(X|Y) \) vary linearly with \( \theta \) as \( \theta \) varies between 0 and 1. The differences \( W_{\Gamma^{\theta}}(X;Y) = W_{\Gamma^{\theta}}(X|Y) - W_{\mathcal{Q}^{\theta}}(X) \) and \( W_{\Pi^{\theta}}(X;Y) = I_{\mathcal{M}^{\theta}}(X|Y) - W_{\Pi^{\theta}}(X|Y) \)
also vary linearly with \( \theta \):

\[
W_{\Gamma^\theta}(X; Y) = (1 - \theta) W_{\Gamma^0}(X; Y) + \theta W_{\Gamma^1}(X; Y),
\]

\[
W_{\Pi^\theta}(X; Y) = (1 - \theta) W_{\Pi^0}(X; Y) + \theta W_{\Pi^1}(X; Y).
\]

E. On the duality between contraction and restriction

If \( \mathcal{M} \) is a convex family of \( \sigma \)-finite measures on \( \mathcal{X} \) then \( I_{\mathcal{M}}(X) \) is a convex functional of \( P_X \) and the convex conjugate functional is defined as \( \Lambda_{\mathcal{M}}(V) = \sup_{m \in \mathcal{M}} \Lambda_m(V) \) for any measurable function \( V(x) \) on \( \mathcal{X} \). We now consider a transition kernel \( \mu(dx|y) \) and show how the convex conjugate of \( I_\mu(X|Y) \) yields that of \( I_{\mathcal{M}}(X) \) where \( \mathcal{M} = \mathcal{K}(\mu) \) is the family of mixtures \( m^\mathcal{B}(dx) = \int_y \mu(dx|y) B(dy) \). The following lemma will be applied.

**Lemma 5.** Suppose \( J^{(2)}(P_{XY}) \) is a convex functional of the joint distribution \( P_{XY} = P_X P_{Y|X} \) and the convex conjugate of the functional \( J^{(2)}(P_{XY}) \) is denoted by \( \Lambda^{(2)}(U) \) for any measurable function \( U(x|y) \) on \( \mathcal{X} \times \mathcal{Y} \). Then \( J^{(1)}(P_X) = \inf_{P_{Y|X}} J^{(2)}(P_X P_{Y|X}) \) is a convex functional of \( P_X \), and the convex conjugate of \( J^{(1)}(P_X) \) is defined for any measurable function \( V(x) \) as \( \Lambda^{(1)}(V) = \Lambda^{(2)}(U_V) \), where \( U_V(x|y) = V(x) \).

**Proof:** The functional \( J^{(1)}(P_X) \) is convex in \( P_X \), since its epigraph is a projection of the epigraph of \( J^{(2)}(P_{XY}) \), which is convex. The convex conjugate of \( J^{(1)}(P_X) \) is given by

\[
\Lambda^{(1)}(V) = \sup_{P_X} [(P_X, V) - J^{(1)}(P_X)]
\]

\[
= \sup_{P_X} \left[ (P_X, V) - \inf_{P_{Y|X}} J^{(2)}(P_X P_{Y|X}) \right]
\]

\[
= \sup_{P_X} \sup_{P_{Y|X}} [(P_X P_{Y|X}, U_V) - J^{(2)}(P_X P_{Y|X})]
\]

\[
= \sup_{P_X} [(P_{XY}, U_V) - J^{(2)}(P_{XY})] = \Lambda^{(2)}(U_V).
\]

This completes the proof of the lemma.

The functional \( J^{(1)}(P_X) = \inf_{P_{Y|X}} J(P_X P_{Y|X}) \) is obtained by contraction from the convex functional \( J^{(2)}(P_{XY}) \), whereas \( \Lambda^{(1)}(V) = \Lambda^{(2)}(U_V) \) is obtained by restriction from the convex conjugate functional \( \Lambda^{(2)}(U) \). It is a general principle of convex duality theory that contraction of a convex functional amounts to restriction of its convex conjugate.

We apply the lemma in the special case when \( J^{(2)}(P_{XY}) \) is the information divergence \( I_\mu(X|Y) \) relative to some transition kernel \( \mu(dx|y) \). For any measurable function \( U(x|y) \) on \( \mathcal{X} \times \mathcal{Y} \) and for any \( y \in \mathcal{Y} \) let \( U^y(x) = U(x|y) \).

**Theorem 23.** The conditional information \( I_\mu(X|Y) \) is convex in the joint distribution \( P(dx dy) \), and its convex conjugate is defined for any measurable function \( U(x|y) \) as

\[
\Lambda^{(2)}(U) = \sup_{U^y} \Lambda_{\mu^y}(U^y).
\]

The functional \( I^{(1)}_{\mu}(X) = \inf_{P(dy|x)} I_\mu(X|Y) \) is equal to \( I_{\mathcal{M}}(X) \) where \( \mathcal{M} = \mathcal{K}(\mu) \), and its convex conjugate is defined for any measurable function \( V(x) \) as

\[
\Lambda^{(2)}(U_V) = \Lambda^{(1)}(V) = \Lambda_{\mathcal{K}(\mu)}(V).
\]

62
Proof: Recall that \( I_\mu(X|Y) \) is the minimum divergence of \( P(dx\,dy) \) relative to the convex set of measures \( \mu(dx|y) \) \( B(dy) \) where \( B(dy) \) varies over all probability distributions on \( \mathcal{Y} \). By Theorem 23 the convex conjugate of \( I_\mu(X|Y) \) is equal to

\[
\Lambda_\mu^{(2)}(U) = \sup_{B(dy)} \log \left( \int \int e^{U(z|y)} \mu(dx|y) B(dy) \right) \\
= \sup_{y \in \mathcal{Y}} \log \left( \int e^{U(y|z)} \mu(y) dx \right) \\
= \sup_{y \in \mathcal{Y}} \Lambda_{\mu^y}(U^y).
\]

The second part of the theorem follows from the first part and Lemma 5. □

Let \( \Pi \) be a \( y \)-convex family of transition kernels \( \mu(dx|y) \), and consider the mixture family \( \mathcal{M} = \mathcal{K}(\Pi) \). If we minimize \( I_\Pi(X|Y) \) over all joint distributions \( P(dx\,dy) \) with marginal \( P(dx) \), then by Theorem 23 we obtain the functional \( I_\mathcal{M}(X) \) which is convex in \( P(dx) \). We compute the convex conjugate of the functional \( I_\Pi(X|Y) \). For any \( y \in \mathcal{Y} \) let \( \Pi^y \) denote the family of measures \( \mu^y(dx) = \mu(dx|y) \) where \( \mu(dx|y) \) ranges over \( \Pi \).

Theorem 24. If \( \Pi \) be a \( y \)-convex family of transition kernels \( \mu(dx|y) \) then \( I_\Pi(X|Y) = \inf_{\mu \in \Pi} I_\mu(X|Y) \) is a convex functional of \( P(dx\,dy) \). The convex conjugate functional is defined for any measurable function \( U(x|y) \) as

\[
\Lambda_\Pi^{(2)}(U) = \sup_{y \in \mathcal{Y}} \Lambda_{\Pi^y}(U^y).
\]

The functional \( I_\Pi^{(1)}(X) = \inf_{P(dy|x)} I_\Pi(X|Y) \) is equal to the convex functional \( I_\mathcal{M}(X) \) where \( \mathcal{M} = \mathcal{K}(\Pi) \), and its convex conjugate is defined for any measurable function \( V(x) \) as

\[
\Lambda_\Pi^{(2)}(U_V) = \Lambda_\Pi^{(1)}(V) = \Lambda_\mathcal{M}(V).
\]

Proof: The functional \( I_\Pi(X|Y) \) is convex in \( P(dx\,dy) \) since it is equal to \( W_\Gamma(X|Y) \) where \( \Gamma \) is the \( y \)-conditional polar of \( \Pi \). The convex conjugate of \( I_\Pi(X|Y) \) is equal to

\[
\Lambda_\Pi^{(2)}(U) = \sup_{P(dx\,dy)} \left[ (P,U) - \inf_{\mu \in \Pi} I_\mu(X|Y) \right] \\
= \sup_{\mu \in \Pi} \sup_{P(dx\,dy)} \left[ (P,U) - I_\mu(X|Y) \right] \\
= \sup_{\mu \in \Pi} \Lambda_\mu^{(2)}(U) \\
= \sup_{\mu \in \Pi} \sup_{y \in \mathcal{Y}} \Lambda_{\mu^y}(U^y) \\
= \sup_{y \in \mathcal{Y}} \Lambda_{\Pi^y}(U^y).
\]

Again, the second part of the theorem follows from the first part and Lemma 5. □

63
VII. ITERATIVE PROCEDURES FOR COMPUTING
LOG-OPTIMUM SELECTIONS IN CONVEX FAMILIES

We formulate an alternating minimization algorithm for computing a log-optimum selection in the convex hull of a parametrized family \( \{ \gamma^y(x) \}_{y \in Y} \). It is related to the algorithm of Blahut [17] for computing points on the rate-distortion curve of a memoryless source, the EM algorithm for computing maximum likelihood estimates of mixture distributions, and the algorithm of Cover [25] for computing log-optimum portfolios. Alternating minimization algorithms have been discussed in great generality by Csiszár and Tusnády [38]. Iussem and Teboulle [65] proposed a primal-dual iterative algorithm for computing the log-optimum \( m^*(dx) \) in the polar \( \mathcal{M}(Q) \) of a finite collection \( Q \).

A. Completely convex hull families

Let \( X \) be a random variable with distribution \( P(dx) \) on a measurable space \( \mathcal{X} \). Let \( Y \) be a second measurable space and let \( \gamma(x|y) \geq 0 \) be a measurable function on the product space \( \mathcal{X} \times \mathcal{Y} \). We consider the convex family \( Q = \mathcal{K}(\gamma) \) of functions of the form

\[
q^B(x) = \int_Y \gamma(x|y) B(dy) = \int \gamma^y(x) B(dy),
\]

where \( \gamma^y(x) = \gamma(x|y) \) and \( B(dy) \) ranges over all probability measures on \( \mathcal{Y} \). Thus \( Q \) is the completely convex hull of the parametrized family \( \{ \gamma^y : y \in Y \} \), or the mixture family \( \mathcal{K}(\Gamma) \) associated with the singleton set \( \Gamma = \{ \gamma \} \).

One may regard \( Y \) as a set of elementary investment opportunities and \( B(dy) \) as a portfolio or betting measure. Suppose an investment of one monetary unit in stock \( y \) will return \( \gamma^y(X) \geq 0 \) units when \( X \) is revealed. If the investor diversifies his capital according to the portfolio \( B(dy) \), then his capital will grow by a factor equal to the weighted average \( q^B(X) = \int \gamma^y(X) B(dy) \). The maximum growth rate of compounded wealth is given by

\[
W_Q(P) = \sup_{B(dy)} E_P \{ \log q^B(X) \}.
\]

To maximize the growth rate of compounded wealth, the investor should diversify according to a portfolio \( B^*(dy) \) such that \( q^*(x) = \int \gamma(x|y) B^*(dy) \) is log-optimum in \( Q \) under \( P \).

Suppose \( W_Q(P) \) is well defined and finite. The maximum growth exponent \( W_Q(P) \) is equal to the minimum information divergence \( I_{\mathcal{M}(Q)}(P) \) relative to measures in the polar family \( \mathcal{M}(Q) \). Clearly, \( \mathcal{M}(Q) \) is the family of measures \( m(dx) \) such that \( \int m(dx) \gamma^y(x) \leq 1 \) for all \( y \in Y \). By specializing Theorem 21 to the case where \( \Gamma \) is the singleton set containing only the function \( \gamma(x|y) \), we see that

\[
W_Q(P) = \sup_{F(dy|x)} [E_P \{ \log \gamma(X|Y) \} - I_{PF}(X; Y)],
\]

(25)

where the supremum is taken over all test channels \( F(dy|x) \) and \( I_{PF}(X; Y) \) is the mutual information when \( X \) and \( Y \) have have joint distribution \( P(dx) F(dy|x) \) on \( \mathcal{X} \times \mathcal{Y} \).

If \( \gamma(x|y) \) is 0-1 valued then each \( \gamma^y \) is the indicator function of a subset \( C^y \subseteq \mathcal{X} \):

\[
x \in C^y \iff \gamma^y(x) = 1.
\]

Let \( C \) denote the family \( \{ C^y \}_{y \in Y} \) and let

\[
H_C(P) = \inf I_{PF}(X; Y),
\]

64
where the infimum is taken over all test channels $F(dy|x)$ such that $X \in C^Y$ or equivalently $\gamma(X|Y) = 1$ almost surely under $P(dx) F(dy|x)$. Since $E\{\log \gamma(X|Y)\} = 0$ for such transition kernels and $E\{\log \gamma(X|Y)\} = -\infty$ otherwise, it follows from (25) that

$$H_c(P) = -W_Q(P)$$

where $Q$ is the completely convex hull of the family $\{\gamma^y : y \in Y\}$. Alternatively,

$$H_c(P) = -I_{M(Q)}(P) = \sup_{m \in M(Q)} E_P \left\{ \log \left( \frac{dm}{dP}(X) \right) \right\},$$

where $M(Q)$ is the family of all measures $m(dx)$ such that $m(C^y) \leq 1$ for all $y \in Y$. In particular, if $X$ is the vertex set of a graph and $C$ is the class of independent subsets of that graph, then $H_c(P)$ is what Körner [73] has called the entropy of the graph. See also Csiszár et al. [37].

**B. Algorithm for computing log-optimum selections**

Let $X$ be a random variable with distribution $P(dx)$ on a measurable space $\mathcal{X}$ and let $Y$ be a second measurable space. Let $\gamma(x|y) \geq 0$ be a measurable function on $\mathcal{X} \times Y$. We describe an iterative procedure for computing a log-optimum $q^*(x)$ in $Q = K(\gamma)$.

For any test channel $F(dy|x)$ we define the marginal distribution

$$B^F(dy) = \int_{\mathcal{X}} P(dx) F(dy|x),$$

and for any probability distribution $B(dy)$ we define the test channel

$$F^B(dy|x) = \frac{\gamma(x|y) B(dy)}{\int_{Y} \gamma(x|y) B(dy)} = \frac{\gamma(x|y) B(dy)}{q^B(x)}.$$

Let $I(PF|m^*\gamma B)$ denote the information of the joint distribution $P(dx)F(dy|x)$ relative to the measure $m^*(dx)\gamma(x|y)B(dy)$ on $\mathcal{X} \times Y$. Obviously,

$$I(PF|m^*\gamma B) = I(PF|m^*\gamma B^F) + I(B^F|B).$$

Let $X$ and $Y$ have joint distribution $P(dx)F(dy|x)$. If $F(dy|x)$ is fixed then $I(PF|m^*\gamma B)$ is minimized when $B(dy)$ is equal to the marginal $B^F(dy)$. The minimum is equal to

$$\inf_{B(dy)} I(PF|m^*\gamma B) = I(PF|m^*\gamma B^F) = I_{\tilde{G}_\gamma}(X|Y) = \tilde{W}_\gamma(X;Y),$$

that is the functional $\tilde{W}_\gamma(X;Y)$ in the special case when $\Gamma$ is the singleton family containing only the function $\gamma(x|y)$. Since $P(dx) = q^*(x)m^*(dx)$, we also have

$$I(PF|m^*\gamma B) = E \left\{ \log \left( \frac{P(dx)F(dy|x)}{m^*(dx)\gamma(X|Y)B(dy)} \right) \right\}$$

$$= E \left\{ \log \left( \frac{q^*(X)F(dy|x)}{\gamma(X|Y)B(dy)} \right) \right\}$$

$$= E \left\{ \log \left( \frac{q^*(X)}{q^B(X)} \right) \right\} + I(PF|PF^B).$$
If $B(dy)$ is fixed then $I(PF|m^*\gamma B)$ is minimized when $F(dy|x)$ is equal to $F^B(dy|x)$, making $I(PF|PF^B) = 0$. The minimum is given by

$$\inf_{F(dy|x)} I(PF|m^*\gamma B) = I(PF^B|m^*\gamma B) = E\left\{ \log \left( \frac{q^*(X)}{q^B(X)} \right) \right\}.$$ 

The value of $I(PF|m^*\gamma B)$ will decrease monotonically if we alternate between minimizing $I(PF|m^*\gamma B)$ over $F(dy|x)$ while $B(dy)$ is fixed, and minimizing $I(PF|m^*\gamma B)$ over $B(dy)$ while $F(dy|x)$ is fixed. This leads to an iterative procedure for computing a log-optimum selection in the convex family $Q$. It is assumed that $W_Q(P)$ is well defined and finite. We start with some distribution $B_0(dy)$ and for $k = 0, 1, 2, \ldots$ we recursively define

$$q_k(x) = q^{B_k}(x) = \int_Y \gamma(x|y) B_k(dy),$$

$$F_k(dy|x) = F^{B_k}(dy|x) = \frac{\gamma(x|y) B_k(dy)}{q_k(x)} = \frac{\gamma(x|y) B_k(dy)}{\int_Y \gamma(x|y) B_k(dy)},$$

$$B_{k+1}(dy) = B^{F_k}(dy) = \int_X P(dx) F_k(dy|x).$$

For all $k \geq 0$ we have

$$I(PF_k|m^*\gamma B_k) \geq I(PF_k|m^*\gamma B_{k+1}) \geq I(PF_{k+1}|m^*\gamma B_{k+1}),$$

so $I(PF_k|m^*\gamma B_k)$ and $I(PF_k|m^*\gamma B_{k+1})$ decrease to the same nonnegative limit. The limit will vanish if $B_0(dy)$ is suitably chosen. If $Y$ is finite then $B_n$ will converge to a log-optimum $B_\infty$. We follow Csiszár [34], who proved convergence of Blahut’s alternating minimization procedure for computing points on the rate-distortion curve of a memoryless source.

**Theorem 25.** Suppose the maximum growth exponent $W_Q(P)$ is well defined and finite, and $B^*(dy)$ is a distribution on $Y$ such that the mixture $q^*(x) = \int \gamma(x|y) B^*(dy)$ is log-optimum in $Q$ under $P(dx)$. If $B_0(dy)$ is a distribution such that $I(B^*|B_0) < \infty$ and if $q_k(x)$ is defined as above, then $E_P\{\log q_k(X)\}$ increases to $W_Q(P) = E_P\{\log q^*(X)\}$ and

$$\log q_k(X) \rightarrow \log q^*(X) \quad \text{in } L^1(P). \quad (27)$$

If $Y$ is a finite set then $B_k$ converges to a limit $B_\infty$ such that $q_\infty(x) = \sum_y \gamma(x|y) B_\infty(y)$ is log-optimum in $Q$.

**Proof:** Consider the test channel $F^*(dy|x) = F^{B^*}(dy|x)$. Clearly $P(dx) F^*(dy|x) = m^*(dx) \gamma(x|y) B^*(dy)$ and $B^*(dy) = B^{F^*}(dy)$ since specialization of (26) yields

$$0 = I(PF^*|m^*\gamma B^*) = I(PF^*|m^*\gamma B^{F^*}) + I(B^{F^*}|B^*)$$

and both terms on the right hand side are nonnegative. The marginals $B^*$ and $B_{k+1}$ are the restrictions of $PF^*$ and $PF_k$ to the $\sigma$-field on $Y$. The divergence between two probability measures can only decrease upon restriction to a smaller $\sigma$-field, so

$$I(PF^*|PF_k) \geq I(B^*|B_{k+1})$$

66
and consequently
\[
0 \leq E \left\{ \log \left( \frac{q^*(X)}{q_k(X)} \right) \right\} = \int P(dx) F^*(dy|x) \log \left( \frac{\gamma(x|y) B^*(dy)}{F^*(dy|x) \gamma(x|y) B_k(dy)} \right) \\
= I(B^*|B_k) - I(PF^*|PF_k) \\
\leq I(B^*|B_k) - I(B^*|B_{k+1}).
\]

Thus \(I(B^*|B_k)\) is monotonically decreasing with \(k\). Since \(I(B^*|B_0)\) is finite by assumption, it follows that \(I(B^*|B_k)\) converges to some finite limit and hence \(E\{\log[q^*(X)/q_k(X)]\}\) decreases to 0. Using Lemma 3, we may conclude that (27) holds. If \(\mathcal{Y}\) is a finite set then some subsequence \(B_{k_i}\) converges to some limiting distribution \(B_\infty\) on \(\mathcal{Y}\). The distribution \(B_\infty\) is log-optimum since \(q_k(x) = \sum_{y \in \mathcal{Y}} \gamma(x|y) B_k(y)\) converges to \(q_\infty(x) = \sum_{y \in \mathcal{Y}} \gamma(x|y) B_\infty(y)\) by continuity and \(\log q_k(X)\) converges to \(q^*(X)\) in \(L^1\). Clearly, \(B_k \to B_\infty\) since \(I(B_\infty|B_k)\) is monotonically decreasing and \(I(B_\infty|B_k)\) vanishes along the subsequence \(k_i\).

Our iterative procedure is related to the algorithm of Cover [25] for computing log-optimum portfolios. Cover combined the two steps \(B_k \mapsto F_k\) and \(F_k \mapsto B_{k+1}\) into a single step \(B_k \mapsto B_{k+1}\). In other words, he examined convergence under repeated application of the transformation which replaces a portfolio \(B(dy)\) by the improved portfolio
\[
B'(dy) = B^{FB}(dy) = \int_{\mathcal{X}} P(dx) F^B(dy|x) = \int_{\mathcal{X}} P(dx) \frac{\gamma(x|y)}{q^B(x)} B(dy).
\]

The advantage in growth exponent of \(B'(dy)\) relative to the given portfolio \(B(dy)\) is bounded as follows: if \(q(x) = q^B(x)\) and \(q'(x) = q^{B'}(x)\) then
\[
E \left\{ \log \left( \frac{q'(X)}{q(X)} \right) \right\} \geq I(B'|B) \geq 0.
\]

Any log-optimum portfolio \(B\) must satisfy the fixed point equation \(B = B'\). Cover assumed that \(\mathcal{Y}\) is finite and proved that \(q_n(X)\) converges to \(q^*(X)\); Csiszár and Tusnády [38] observed that the portfolios \(B_n\) themselves converge to a log-optimum portfolio \(B^*\).

Let \(\tilde{W}^{(k)}(X;Y)\) denote the functional \(\tilde{W}_n(X;Y)\) when \(X\) and \(Y\) have joint distribution \(P(dx) F_k(dy|x)\). Specializing (26), we see that
\[
I(PF_k|m^* \gamma B_k) = I(PF_k|m^* \gamma B_{k+1}) + I(B_{k+1}|B_k)
\]
or equivalently
\[
E \left\{ \log \left( \frac{q^*(X)}{q_k(X)} \right) \right\} = \tilde{W}^{(k)}(X;Y) + I(B_{k+1}|B_n).
\]

The left hand side vanishes in the limit as \(k \to \infty\) by Theorem 25. Thus \(I(B_{k+1}|B_k) \to 0\) and the minimum divergence \(\tilde{W}^{(k)}(X;Y)\) of \(P(dx) F_k(dy|x)\) relative to the convex set of distributions \(m^*(dx) \gamma(x|y) B(dy)\) decreases to 0 as \(k \to \infty\).

C. Computing points on the rate-distortion curve

We consider a source alphabet \(\mathcal{X}\), a reproduction alphabet \(\mathcal{Y}\), and a measurable distortion function \(\rho(x,y)\) on \(\mathcal{X} \times \mathcal{Y}\). The rate-distortion function of a random variable \(X\) with distribution \(P(dx)\) on \(\mathcal{X}\) is defined as
\[
R(D) = \inf_{F: E_{PP}(\rho(X,Y)) \leq D} I_{PP}(X;Y),
\]

67
where the minimum mutual information is taken over all test channels $F(dy|x)$ such that the expected distortion $E_{PF}\{\rho(X,Y)\}$ under the joint distribution $P(dx)F(dy|x)$ does not exceed $D$. The function $R(D)$ is convex and hence continuous on the interior of the interval on which it is finite. The lower semicontinuous regularization is given by
\[
R(D+) = \lim_{\epsilon \downarrow 0} R(D + \epsilon).
\]

For $s \leq 0$ let $\gamma_s(x|y) = e^{s\rho(x,y)}$ and let $W(s) = W_{Q_s}(P)$ denote the maximum growth exponent of selections in the convex family $Q_s$ of functions of the form
\[
a_s^P(x) = \int_Y \gamma_s(x|y) B(dy) = \int_Y e^{s\rho(x,y)} B(dy),
\]
where $B(dy)$ varies over all probability distributions on $Y$. The following result is well known, cf. Gallager [52], Berger [14], Blahut [18], Csiszár [33], Csiszár and Körner [36].

**Theorem 26.** If the distortion function $\rho(x,y)$ is bounded below on $X \times Y$ then
\[
W(s) = \sup_{D} [sD - R(D)], \quad s \leq 0,
\]
and consequently, by convex duality,
\[
R(D+) = \sup_{s \leq 0} [sD - W(s)].
\]

**Proof:** Applying (25) with $\gamma_s(x|y) = e^{s\rho(x,y)}$ and $Q_s = K(\gamma_s)$, we see that
\[
W(s) = \sup_{F(dy|x)} [sE_{PF}\{\rho(X,Y)\} - I_{PF}(X;Y)],
\]
where the supremum is taken over all test channels $F(dy|x)$. One may take the supremum first over all $F(dy|x)$ with $E_{PF}\{\rho(X,Y)\} = \Delta$ and then over $-\infty < \Delta \leq \infty$. Thus
\[
W(s) = \sup_{-\infty < \Delta \leq \infty} [s\Delta - R^c(\Delta)],
\]
where
\[
R^c(\Delta) = \inf_{F: E_{PF}\{\rho(X,Y)\} = \Delta} I_{PF}(X;Y).
\]
The value $\Delta = -\infty$ may be excluded since $\rho(x,y)$ is bounded below, and if $s < 0$ then the value $\Delta = +\infty$ may be excluded as well since the quantity to be maximized tends to $-\infty$ as $\Delta \to \infty$. Since $R(D) = \inf_{\Delta \leq D} R^c(\Delta)$, we have $R(D) \leq R^c(D)$ and
\[
W(s) = \sup_{D} [sD - R^c(D)] \leq \sup_{D} [sD - R(D)].
\]
On the other hand, if $s \leq 0$ and $\Delta \leq D$ then $sD - R^c(\Delta) \leq s\Delta - R^c(\Delta)$, hence
\[
sD - R(D) = \sup_{\Delta \leq D} [sD - R^c(\Delta)] \leq \sup_{\Delta \leq D} [s\Delta - R^c(\Delta)].
\]
Taking the supremum over all \( D \) proves that
\[
\sup_D [sD - R(D)] \leq \sup \Delta [s\Delta - R^e(\Delta)] = W(s).
\]
Thus \( \sup_D [sD - R(D)] = \sup \Delta [s\Delta - R(\Delta)] = W(s) \). The theorem follows. \( \square \)

The alternating minimization algorithm of Section B, in the special case when \( \gamma(x|y) \) is equal to \( \gamma_s(x|y) = e^{s\rho(x,y)} \), reduces to the algorithm of Blahut [17] [18] for computing a point \((D_s, R(D_s))\) where the rate-distortion curve \( R(D) \) has slope \( s \). We start with a suitable distribution \( B_0(dy) \) and we minimize \( I(PP|P\gamma_sB) \) over \( F(dy|x) \) while \( B(dy) \) is fixed, then over \( B(dy) \) while \( F(dy|x) \) is fixed, and repeat. The result is a sequence of transition kernels \( F_k(dy|x) \) and distributions \( B_{k+1}(dy) \) so that
\[
\inf_{F(dy|x)} I(PP|P\gamma_sB_k) = I(PP_k|P\gamma_sB_k) = -E_P\{\log \gamma^B_k(X)\},
\]
\[
\inf_{B(dy)} I(PP_k|P\gamma_sB) = I(PP_k|P\gamma_sB_{k+1}) = I_{PF_k}(X;Y) = - s E_{P}\{\rho(X;Y)\}.
\]
Both sequences decrease monotonically to \(-W(s)\). In fact,
\[
-W(s) = \inf_{F(dy|x)} [I_{PF}(X;Y) - sE_{P}\{\rho(X;Y)\}] = - \sup_{B(dy)} E_P \{\log \gamma^B(X)\}.
\]
Conversely, the alternating minimization algorithm of Section B is the Blahut algorithm when \( s = -1 \) and the distortion function is \( \rho(x,y) = -\log \gamma(x|y) \).

If \( s \leq 0 \) then \( W(s) \) is minus the vertical intercept at \( D = 0 \) of the line with slope \( s \) that supports the epigraph of the convex nonincreasing function \( R(D) \). The function \( W(s) \) for \( s \leq 0 \) uniquely determines the function \( R(D) \) except at the boundary point
\[
D_{\min} = E\{\inf_{y \in Y} \rho(X, y)\}.
\]
In general \( R(D_{\min}) \) may exceed \( R(D_{\min}+) \), but equality holds if \( X \) and \( Y \) are finite sets.

Suppose \( \rho(x,y) \geq 0 \) and for all \( x \in X \) there exists some \( y \in Y \) such that \( \rho(x,y) = 0 \), so that \( D_{\min} = 0 \). An interesting interpretation for \( R(0) \), due to Csiszár et al. [37], can be found by considering the \( \{0,1\} \)-valued function
\[
\gamma_{-\infty}(x|y) = \lim_{s \to -\infty} e^{s\rho(x,y)} = \begin{cases} 1 & \text{if } \rho(x,y) = 0, \\ 0 & \text{if } \rho(x,y) > 0. \end{cases}
\]
It can be shown that \( -W_{Q_s}(P) \to R(0+) \) as \( s \to -\infty \) and that
\[
R(0) = -W_{Q_{-\infty}}(P) = H_C(P),
\]
where \( C = \{C^Y\}_{Y \in Y} \) is the family of subsets of \( X \) of the form
\[
C^Y = \{x \in X : \gamma_{-\infty}(x|y) = 1\} = \{x \in X : \rho(x,y) = 0\}.
\]
Recall that \( P(dx) = q^*_s(x)m^*_s(dx) \) where \( q^*_s(x) \) and \( m^*_s(dx) \) are log-optimum in \( Q_s \) and the polar family \( M(Q_s) \); respectively. The transition kernel \( \tilde{G}^*_s(dy|x) = m^*_s(dx) \gamma_s(x|y) \) is
subnormalized, and the probability measures $B(dy)$ such that $q_s^B(x) = \int \gamma_s(x | y) B(dy)$ is log-optimum in $Q_s$ are exactly those for which the marginal

$$\int \tilde{G}_s^*(dx | y) B(dy) = m_s^*(dx) q_s^B(x)$$

is equal to the true distribution $P(dx)$. Thus the set of distributions $B(dy)$ such that $q_s^B(x)$ is log-optimum in $Q_s$ is the intersection of the simplex of all probability distributions on $Y$ with the affine space of distributions that satisfy the linear constraint $q_s^B(X) = q_s^*(X)$ or

$$\int y \gamma_s(X|y) B(dy) = \frac{dP}{dm_s^*(X)}(X)$$

almost surely.

Each distribution $B_s^*(dy)$ satisfying this constraint defines a point on the rate-distortion curve, namely the point $(D_s, R(D_s))$ where $D_s = E_{PF_s^*} \{\rho(X, Y)\}$ and $R(D_s) = I_{PF_s^*} (X; Y)$ are the expected distortion and the mutual information when $X$ and $Y$ have joint distribution $P(dx) F_s^*(dy | x) = m_s^*(dx) \gamma_s(x | y) B_s^*(dy)$. These points $(D_s, R(D_s))$ are exactly those that belong to the supporting line with slope $s$, i.e. they satisfy the equation

$$R(D_s) = sD_s - W(s).$$

Indeed, the marginals of $m_s^*(dx) \gamma_s(x | y) B_s^*(dy)$ are equal to $P(dx)$ and $B_s^*(dy)$, so the mutual information is given by

$$R(D_s) = I_{PF_s^*} (X; Y) = E_{PF_s^*} \left\{ \log \left( \frac{m_s^*(dX) \gamma_s(X|Y) B_s^*(dY)}{P(dX) B_s^*(dY)} \right) \right\}$$

$$= E_{PF_s^*} \left\{ \log \left( \frac{\gamma_s(X|Y)}{q_s^*(X)} \right) \right\}$$

$$= s E_{PF_s^*} \{\rho(X, Y)\} - E \{ \log q_s^*(X) \}$$

$$= sD_s - W(s).$$

If the supporting line with slope $s$ contains two points $(R^0, D^0)$ and $(R^1, D^1)$ then the supporting line contains the entire line segment connecting these points. The convex conjugate $W(s) = \sup_y [sD - R(D)]$ then has a corner point at $s$: the graph of $W(s)$ is supported at $s$ by lines whose slope varies between $D^0$ and $D^1$.

For any test channel $F(dy|x)$ one may consider the point $(D, R)$ where $D = E_{PF} \{\rho(X, Y)\}$ is the expected distortion and $R = I_{PF} (X; Y)$ is the mutual information when $X$ and $Y$ have joint distribution $P(dx) F(dy|x)$. This point $(D, R)$ belongs to the epigraph of the function $R(D)$. The amount by which the supporting line with slope $s$ must be shifted in the vertical direction so that it passes through the point $(D, R)$ is equal to $\tilde{W}_\gamma_s (X; Y)$ since

$$\tilde{W}_\gamma_s (X; Y) = I(PF|m_s^* \gamma B^F) = E \left\{ \log \left( \frac{q_s^*(X)}{\gamma_s(X,Y)} \frac{F(dY|X)}{B^F(dY)} \right) \right\}$$

$$= I(X; Y) - s E \{\rho(X, Y)\} + E \{ \log q_s^*(X) \}$$

$$= R - sD + W(s)$$

and $R - sD + W(s)$ is exactly the vertical displacement between the point $(D, R)$ and the supporting line with slope $s$. In Blahut’s alternating minimization algorithm, the vertical displacement $\tilde{W}_{\gamma_s}^{(k)} (X; Y) = I(PF_k|m_s^* \gamma B_{k+1})$ decreases monotonically to zero.
Recall the identity $I(X; Y) = W_{\gamma}(X; Y) + \tilde{W}_{\gamma}(X; Y)$. Although $\tilde{W}_{\gamma}(X; Y) \geq 0$, it is not always true that $W_{\gamma}(X; Y) \geq 0$. From the graphical interpretation of $\tilde{W}_{\gamma}(X; Y)$, we see that $W_{\gamma}(X; Y)$ is negative if the supporting line with slope $s$ intersects the horizontal axis at a distortion level less than $D = E\{\rho(X, Y)\}$.

D. The EM algorithm

Let $X$ be a random variable with distribution $P(dx)$ on a measurable space $\mathcal{X}$. Let $\mu(dx|y)$ be a transition kernel from $\mathcal{Y}$ to $\mathcal{X}$ and let $\mathcal{M}$ denote the family of mixtures

$$m^B(dx) = \int_{\mathcal{Y}} \mu(dx|y) B(dy) = \int_{\mathcal{Y}} \mu(dy) B(dy),$$

where $\mu(dy) = \mu(dx|y)$ and where $B(dy)$ ranges over all probability measures on $\mathcal{Y}$. Thus $\mathcal{M} = \mathcal{K}(\mu)$ is the completely convex hull of the parametrized family $\{\mu^y : y \in \mathcal{Y}\}$. We wish to find a distribution $B^* (dy)$ on $\mathcal{Y}$ such that the mixture $m^* (dx) = \int \mu(dx|y) B^* (dy)$ attains the minimum divergence $I_{\mathcal{M}}(P) = \inf_{B(dy)} I(P|m^B)$.

It turns out that minimizing the information of $P(dx)$ relative to measures in the mixture family $\mathcal{M} = \mathcal{K}(\mu)$ is equivalent to maximizing the growth exponent of functions in the mixture family $\mathcal{Q} = \mathcal{K}(\gamma)$, for some function $\gamma(x|y)$. Indeed, let $\lambda(dx)$ be a $\sigma$-finite reference measure that dominates the family $\{\mu^y : y \in \mathcal{Y}\}$. The existence of such a dominating measure can be assumed without loss of generality because the probability measure $P(dx)$ is fixed and each $\mu^y(dx)$ may be replaced by its $P$-dominated part. Let $\gamma(x|y) = \gamma^y(x)$ denote the density of $\mu(dx|y) = \mu^y(dx)$ relative to $\lambda(dx)$. Then the density $dm^B/d\lambda(x)$ is equal to $q^B(x) = \int \gamma^y(x) B(dy)$, and the family $\mathcal{Q}$ of such densities $q^B(x)$ is convex. Maximizing the growth exponent $E_P\{\log q^B(X)\}$ is equivalent to minimizing the information $E\{\log [dP/dm^B(X)]\}$ since the sum is a constant independent of $B$:

$$E_P\{\log q^B(X)\} + E_P \left\{ \log \left( \frac{dP}{dm^B(X)} \right) \right\} = I_{\lambda}(P).$$

Thus $W_{\mathcal{Q}}(P) + I_{\mathcal{M}}(P) = I_{\lambda}(P)$, and the measures $B^* (dy)$ such that $m^* (dx) = \int \mu(dx|y) B^* (dy)$ is log-optimum in $\mathcal{M} = \mathcal{K}(\mu)$ are exactly those such that $q^* (x) = \int \gamma(x|y) B^* (dy)$ is log-optimum in $\mathcal{Q} = \mathcal{K}(\gamma)$. Also, a test channel $F^*(dy|x)$ attains $I_{\mathcal{M}}(P) = \inf_{F(dy|x)} I_{\mu}(X|Y)$ iff it attains

$$W_{\mathcal{Q}}(P) = \sup_{F(dy|x)} \left[ E\{\log \gamma(X|Y)\} - I(X; Y) \right]$$

$$= I_{\lambda}(X) - \inf_{F(dy|x)} I_{\mu}(X|Y)$$

$$= I_{\lambda}(X) - I_{\mathcal{M}}(X).$$

The alternating minimization algorithm starts with a distribution $B_0(dy)$ and produces distributions $B_k(dy)$ for successive $k = 1, 2, \ldots$ such that $q_k(x) = \int \gamma(x|y) B_k(dy)$ is log-optimum in $\mathcal{Q}$ in the limit as $k \to \infty$. In terms of $m_k(dx) = q_k(x) \lambda(dx)$, the alternating minimization procedure can be formulated as follows:

$$m_k(dx) = \int_{\mathcal{Y}} \mu(dx|y) B_k(dy),$$

$$F_k(dy|x) = \frac{\mu(dx|y) B_k(dy)}{m_k(dx)} = \frac{\mu(dx|y) B_k(dy)}{\int_{\mathcal{Y}} \mu(dx|y) B_k(dy)},$$

71
\[ B_{k+1}(dy) = \int_X P(dx) F_k(dy|x). \]

The two steps \( B_k(dy) \rightarrow F_k(dy|x) \rightarrow B_{k+1}(dy) \) can be combined into the single step

\[ B_{n+1}(dy) = \int_X P(dx) \frac{\mu(dx|y)}{\int y \mu(dx|y) B_k(dy)} B_k(dy). \]

Suppose \( B^*(dy) \) is a distribution on \( Y \) such that the mixture \( m^*(dx) = \int_Y \mu(dx|y) B^*(dy) \) attains the minimum divergence \( I_{\mathcal{M}}(P) = \inf_{B(dy)} I_{m^B}(P) \). If the starting point \( B_0(dy) \) is chosen so that \( I(B^*|B_0) < \infty \), then \( I_{m_k}(P) \rightarrow I_{\mathcal{M}}(P) \) as \( k \rightarrow \infty \). In fact,

\[ \log \left( \frac{dP}{dm_k}(X) \right) \rightarrow \log \left( \frac{dP}{dm^*}(X) \right) \quad \text{in } L^1(P). \]

In statistical applications, the true distribution \( P(dx) \) is unknown and replaced by the empirical distribution based on a finite number of independent samples of \( X \). Also, \( \mu(dx|y) \) is normalized and is interpreted as the conditional distribution of \( X \) given the latent random variable \( Y \). The alternating minimization procedure for computing \( B^*(dy) \) is nothing but the EM algorithm for computing maximum likelihood estimates of mixing distributions from incomplete data. The joint distribution \( \mu(dx|y) B^*(dy) \) is then the maximum likelihood estimate of the distribution of the complete data \( (X,Y) \).

The EM algorithm was discussed by Dempster, Laird and Rubin [39] and earlier in special cases by many other authors. See also Redner and Walker [105]. Vardi and Lee [123] have proposed the EM algorithm as a general-purpose method for solving linear inverse problems with positivity constraints (so-called LININPOSS problems). They emphasize engineering applications such as signal recovery or deconvolution, tomographic reconstruction, motion deblurring and image restoration. Lindsay [87] discusses mixture models from a geometrical point of view.

E. Alternating minimization in a more general setting

We generalize the alternating minimization algorithms in Sections B and E. For further generalizations see Csiszár and Tusnády [38].

Let \( \Gamma \) be a \( y \)-convex family of measurable functions \( \gamma(x|y) \geq 0 \) on \( X \times Y \), and suppose \( q^*(x) = \int \gamma^*(x|y) B^*(dy) \) is log-optimum in the mixture family \( \mathcal{Q} = \mathcal{K}(\Gamma) \) under some distribution \( P(dx) \). Then \( m^*(dx) = P(dx)/q^*(x) \) is log-optimum in the polar family \( \mathcal{M}(\mathcal{Q}) \). The divergence \( I(PF|m^*\gamma B) \) will decrease monotonically if we alternate between minimizing \( I(PF|m^*\gamma B) \) over \( F(dy|x) \) while \( B(dy) \) and \( \gamma(x|y) \) are fixed, and minimizing \( I(PF|m^*\gamma B) \) over \( B(dy) \) and \( \gamma(x|y) \) while \( F(dy|x) \) is fixed. Suppose we start with some function \( \gamma_0(x|y) \) in \( \Gamma \) and some distribution \( B_0(dy) \) on \( Y \). For successive \( k = 0, 1, 2, \ldots \) let

\[ q_k(x) = \int_Y \gamma_k(x|y) B_k(dy), \]

\[ F_k(dy|x) = \frac{\gamma_k(x|y) B_k(dy)}{q_k(x)}, \]

let \( B_{k+1}(dy) = \int_X P(dx) F_k(dy|x) \), and let \( \gamma_{k+1}(x|y) \) be conditionally log-optimum given \( y \) in \( \Gamma \) under \( P(dx) F_k(dy|x) \). As before, \( I(PF_k|m^*\gamma_k B_k) = E[\log[q^*(X)/q_k(X)]] \) and
\[ I(P F_k | m^* \gamma_{k+1} B_{k+1}) = \tilde{W}_r^{(k)}(X; Y) \] decrease to the same nonnegative limit. By Theorem 3 of Csiszár and Tusnády [38], the limit is equal to the minimum divergence between two sets of measures on \( \mathcal{X} \times \mathcal{Y} \), and the limit will vanish if \( \gamma_0(x|y) \) and \( B_0(dy) \) are suitably chosen. Indeed, consider the set \( \mathcal{P} \) of all joint distributions \( P(dx) F(dy|x) \) with marginal \( P(dx) \) and the set \( \mathcal{P}_r \) of subprobability measures \( m^*(dx) \gamma(x|y) B(dy) \) on \( \mathcal{X} \times \mathcal{Y} \). Then \( \mathcal{P} \) and \( \mathcal{P}_r \) are convex and the limit in question is equal to the minimum divergence \( I(\mathcal{P}_0 | \mathcal{P}_r) \) where \( \mathcal{P}_0 \subseteq \mathcal{P} \) is the subset of distributions \( P(dx) F(dy|x) \) such that \( I(P F | m^* \gamma_k B_k) < \infty \) for some \( k \). For suitable \( \gamma_0(x|y) \) and \( B_0(dy) \), the limit is equal to \( I(\mathcal{P} | \mathcal{P}_r) = 0 \) and

\[
\log q_k(X) \rightarrow \log q^*(X) \quad \text{in} \ L^1(P).
\]

Dual results hold if \( \Pi \) is a \( y \)-convex family of transition kernels \( \mu(dx|y) \geq 0 \) from \( \mathcal{Y} \) to \( \mathcal{X} \). Suppose \( m^*(dx) = \int \mu^*(dx|y) B^*(dy) \) is log-optimum in the mixture family \( \mathcal{M} = \mathcal{K}(\Pi) \), and \( q^*(x) = dP/dm^*(x) \) is log-optimum in the polar family \( \mathcal{Q}(\mathcal{M}) \). The divergence \( I(P F | q^* \mu B) \) will decrease monotonically if we alternate between minimizing \( I(P F | q^* \mu B) \) over \( F(dy|x) \) while \( B(dy) \) and \( \mu(x|y) \) are fixed and minimizing \( I(P F | q^* \mu B) \) over \( B(dy) \) and \( \mu(dx|y) \) while \( F(dy|x) \) is fixed. Suppose we start with some transition kernel \( \mu_0(dx|y) \) in \( \Pi \) and some distribution \( B_0(dy) \) on \( \mathcal{Y} \). For successive \( k = 0, 1, 2, \ldots \) let

\[
m_k(dx) = \int \mu_k(dx|y) B_k(dy),
\]

\[
F_k(dy|x) = \frac{\mu_k(dx|y) B_k(dy)}{m_k(dx)} = \frac{\mu_k(dx|y) B_k(dy)}{\int \mu_k(dx|y) B_k(dy)},
\]

let \( B_{k+1}(dy) = \int_{\mathcal{X}} P(dx) F_k(dy|x) \), and let \( \mu_{k+1}(dx|y) \) be conditionally log-optimum given \( y \) in \( \Pi \) under \( P(dx) F_k(dy|x) \). If \( B_0(dy) \) and \( \mu_0(dx|y) \) are suitably chosen then \( I(P F_k | q^* \mu_k B_k) = E[\log [d m^*/d m_k(X)]] \) and \( I(P F_k | q^* \gamma_{k+1} B_{k+1}) = \tilde{W}_r^{(k)}(X; Y) \) will decrease to zero and

\[
\log \left( \frac{dP}{d m_k}(X) \right) \rightarrow \log \left( \frac{dP}{d m^*}(X) \right) \quad \text{in} \ L^1(P).
\]
VIII. OPTIMAL SPECTRA

A. Maximizing Determinants

An interesting example of a log-optimum selection problem is maximizing $W(S) = \log \det S$ when $S$ ranges over some convex set $\mathcal{S}$ of symmetric positive semidefinite $n \times n$ matrices. It is well known that $\log \det S = \text{tr} \log S$ is a concave function of $S$ and that

$$\delta[\log \det S] = \delta[\text{tr} \log S] = \text{tr}[\delta \log S] = \text{tr}[S^{-1} \delta S].$$

Let $\Sigma$ denote a matrix with maximum determinant in the class $\mathcal{S}$. The directional derivative of $W(S) = \log \det S$ as we move away from $S_0 = \Sigma$ with constant speed along the line segment $S_\epsilon = (1 - \epsilon)S_0 + \epsilon S_1$ toward $S_1 = S$ is given by

$$\lim_{\epsilon \searrow 0} \frac{W(S_\epsilon) - W(S_0)}{\epsilon} = \text{tr}[\Sigma^{-1} (S - \Sigma)] = \text{tr}[(S - \Sigma)\Sigma^{-1}].$$

The Karush-Kuhn-Tucker (KKT) conditions for $\Sigma$ assert that $\text{tr}[S\Sigma^{-1}] \leq n$ for all $S \in \mathcal{S}$, or equivalently that $\Sigma^{-1}$ is orthogonal to all differences $S - \Sigma$:

$$\text{tr}[(S - \Sigma)\Sigma^{-1}] \leq 0, \quad \text{for all } S \in \mathcal{S}.$$

Maximizing the determinant amounts to maximizing the geometric mean of the eigenvalues, since $\det S = \prod_{1 \leq k \leq n} \lambda_k(S)$ is the product of the eigenvalues. If $E_k\{\cdot\}$ denotes expectation when $k$ is random and uniformly distributed over the range $1 \leq k \leq n$, then

$$\frac{1}{n} \log \det S = \frac{1}{n} \sum_{1 \leq k \leq n} \log \lambda_k(S) = E_k\{\log \lambda_k(S)\}.$$

If all matrices in $\mathcal{S}$ commute then they are simultaneously diagonalized by a common system of eigenvectors, the eigenvalues $\lambda_k(S)$ are affine functions of $S$, and the spectrum of the matrix $\Sigma$ with maximum determinant is characterized by the KKT conditions

$$\frac{1}{n} \sum_{1 \leq k \leq n} \frac{\lambda_k(S)}{\lambda_k(\Sigma)} = E_k\left\{\frac{\lambda_k(S)}{\lambda_k(\Sigma)}\right\} \leq 1, \quad \text{for all } S \in \mathcal{S}.$$

The matrix $\Sigma$ with maximum determinant may be called the maximum entropy matrix in the class $\mathcal{S}$, because maximizing $\log \det S$ is equivalent to maximizing the differential entropy $h(X)$ of a gaussian random vector $X$ with mean zero and covariance matrix $S$:

$$h(X) = \frac{n}{2} \log (2\pi e) + \frac{1}{2} \log \det S.$$

Maximizing entropy subject to convex constraints plays an important role in the theory of positive definite completions of partially specified symmetric matrices. Suppose $\mathcal{S}$ is the convex family of all symmetric (or hermitian) positive semidefinite matrices for which certain entries $S_{kl}$ take on specified values. If the entry $S_{rs}$ is not specified and some matrix in $\mathcal{S}$ is strictly positive definite, then the entry $\Sigma^{-1}_{rs}$ of the inverse $\Sigma^{-1}$ must vanish and by symmetry also $\Sigma^{-1}_{sr}$ must vanish. Indeed, one may choose variations $S - \Sigma$ that are positive or negative multiples of the matrix $E^{(r,s)}$ with a single nonzero entry in position $(r, s)$ and conclude from the KKT conditions (28) that

$$\text{tr}[E^{(r,s)}\Sigma^{-1}] = \sum_{k,l} E^{(r,s)}_{kl}\Sigma^{-1}_{lk} = \Sigma^{-1}_{sr} = 0.$$
If the entries of $S$ are specified only in a narrow band around the main diagonal, then $\Sigma^{-1}$ will be a sparse matrix with nonzero entries in that narrow band only. If the locations of the specified entries form the edges of a chordal graph, then the determinant can be maximized in stepwise fashion (by filling in the unspecified items one at a time) and there is an explicit formula for the maximum determinant which generalizes Hadamard's inequality. For more information about positive definite completions see Johnson [68].

Another log-optimum selection problem is computing the capacity of a channel with additive gaussian noise. The capacity of a memoryless channel is the maximum mutual information between the input $X$ and the output $Y$ subject to certain constraints on the input distribution. We assume the channel produces the output vector $Y$ by adding independent gaussian noise $N$ to the input vector $X$:

$$Y = X + N.$$

Suppose the noise vector $N$ has zero mean and positive definite covariance matrix $S_N$, and the input covariance matrix $S_X$ is constrained to some convex set $\mathcal{S}_X$. Maximizing the mutual information $I(X;Y)$ is equivalent to maximizing the output entropy $h(Y)$ because $I(X;Y) = h(Y) - h(Z)$. It suffices to maximize over gaussian input distributions with covariance matrices in $\mathcal{S}_X$, because $S_Y = S_X + S_N$ will remain unchanged while $h(Y)$ can only increase when a non-gaussian input signal vector $X$ is replaced by a gaussian vector with the same covariance matrix $S_X$. Thus maximizing the mutual information $I(X;Y)$ or the output entropy $h(Y)$ amounts to maximizing log det $S_Y$ or maximizing

$$\log \det[S_Y S_N^{-1}] = \log \det[I + S_X S_N^{-1}] = \log \det[I + S_X^{-1} S_N].$$

The optimal input covariance matrix $\Sigma_X$ is characterized by the KKT conditions

$$\text{tr}[(S_X - \Sigma_X)(S_N + \Sigma_X)^{-1}] \leq 0, \quad \text{for all } S_X \in \mathcal{S}_X.$$

If $S_X$ is defined by the input power constraint $\text{tr}[S_X] \leq P$, then $\Sigma_X$ can be obtained by diagonalizing the matrix $S_N$ (or by decomposing the channel into parallel independent subchannels) and applying the classical water pouring technique to its spectrum. The eigenvectors of $\Sigma_X$ are those of $S_Z$ and the eigenvalues satisfy the KKT conditions

$$\sum_k \frac{\lambda_k(S_X) - \lambda_k(S_N)}{\lambda_k(S_N) + \lambda_k(S_X)} \leq 0, \quad \text{for all } S_X \in \mathcal{S}_X.$$

It should be impossible to move input power $\lambda_k(\Sigma_X)$ from modes $k$ where the output variance $\lambda_k(S_N) + \lambda_k(\Sigma_X)$ is high to other modes where the output variance is low. The water pouring solution minimizes the maximum value of the output variance $\lambda_k(S_N) + \lambda_k(\Sigma_X)$ over all modes $k$ at which input power $\lambda_k(\Sigma_X) > 0$ is present.

**B. Maximizing the exponential growth rate of determinants**

Let $\{S_i\}_{-\infty < i < \infty}$ be a two-sided infinite sequence of real numbers such that for every integer $n \geq 1$, the following Toeplitz matrix is symmetric and positive definite:

$$S^n = \begin{pmatrix}
S_0 & S_{-1} & \cdots & S_{-(n-1)} \\
S_1 & S_0 & \cdots & S_{-(n-2)} \\
\vdots & \vdots & \ddots & \vdots \\
S_{n-1} & S_{n-2} & \cdots & S_0
\end{pmatrix}.$$
The Fourier transform of the sequence $\{S_t\}$ is the nonnegative periodic function
\[
S(e^{i\theta}) = \sum_{-\infty < l < \infty} S_l e^{il\theta}.
\]

Szegö proved that under certain conditions on $S(e^{i\theta})$, $\det S^n$ grows geometrically with limiting growth factor equal to the geometric mean of the spectrum $S(e^{i\theta})$. In fact,
\[
\lim_n [\det S^n]^{1/n} = \exp \left( \int_{-\pi}^{\pi} \log S(e^{i\theta}) \frac{d\theta}{2\pi} \right).
\]
The limit can be interpreted as the mean squared prediction error for a stationary gaussian process with spectrum $S(e^{i\theta})$. See Grenander and Szegö [60].

Let $\{\lambda_k(S^n)\}_{1 \leq k \leq n}$ denote the spectrum of the matrix $S^n$ and observe that $\det S^n$ is the product of the eigenvalues. Suppose we pick an eigenvalue at random by selecting its index $k$ according to a uniform distribution on the range $1 \leq k \leq n$. In the limit as $n \to \infty$, $\lambda_k(S^n)$ is asymptotically distributed like the function $S(e^{i\theta})$ of a random variable $\Theta$ that is uniformly distributed according to normalized Lebesgue measure $d\theta/(2\pi)$ on the circle $\theta \in [-\pi, \pi)$. Formally, the random variable $\lambda_k(S^n)$ (where $k$ is uniformly distributed over $1 \leq k \leq n$) converges in distribution to $S(e^{i\theta})$ as $n \to \infty$. The expected value
\[
E_k \{\log \lambda_k(S^n)\} = \frac{1}{n} \sum_{1 \leq k \leq n} \log \lambda_k(S^n) = \frac{1}{n} \log \det S^n
\]
converges to the expected value
\[
E_\Theta \{\log S(e^{i\theta})\} = \int_{-\pi}^{\pi} \log S(e^{i\theta}) \frac{d\theta}{2\pi}
\]
and hence $\det S^n$ grows exponentially fast with limiting rate $E_\Theta \{\log S(e^{i\theta})\}$.

The sequence $\{S_t\}_{-\infty < t < \infty}$ is the autocorrelation function and its Fourier transform $S(e^{i\theta}) = \sum_l S_l e^{il\theta}$ is the power spectral density of a stationary gaussian process $\{X_t\}$ with mean $E\{X_t\} = 0$ and covariances
\[
E\{X_t X_{t+l}\} = S_l.
\]

This gaussian process has maximum entropy rate among all stationary processes with the same spectrum. The entropy rate of $\{X_t\}$ is defined as the limiting entropy per symbol and in the real-valued gaussian case is given by Kolmogorov’s formula
\[
\lim \frac{1}{n} h(X^n) = \frac{1}{2} \log(2\pi e) + \frac{1}{2} \int_{-\pi}^{\pi} \log S(e^{i\theta}) \frac{d\theta}{2\pi}.
\]
Maximizing the entropy rate is equivalent to maximizing the exponential growth rate of $\det S^n$ because
\[
\frac{1}{n} h(X^n) = \frac{1}{2} \log(2\pi e) + \frac{1}{2n} \log \det S^n.
\]
Suppose the spectrum is constrained to some convex set $S$. The spectrum $\Sigma(e^{i\theta})$ in $S$ with maximum entropy rate is the log-optimum spectrum that satisfies the KKT conditions
\[
E_\Theta \left\{ \frac{S(e^{i\theta})}{\Sigma(e^{i\theta})} \right\} = \int_{-\pi}^{\pi} \frac{S(e^{i\theta})}{\Sigma(e^{i\theta})} \frac{d\theta}{2\pi} \leq 1, \quad \text{for all } S(e^{i\theta}) \text{ in } S.
\]
Burg studied the problem of maximizing the entropy rate of a stationary process \( \{X_t\} \) when the correlation coefficients \( E\{X_t X_{t+l}\} \) take on specified values \( S_l = S_{-l} \) for \( 0 \leq l < K \). The covariance matrix of a segment of length \( n \) is equal to the Toeplitz matrix \( S^n \) and its entries \( S_{kl} = E\{X_k X_l\} \) take on specified values \( S_{kl} = S_{l-k} \) in the band \( |l - k| \leq K \). The process with maximum entropy rate subject to these correlation constraints is an autoregressive Gauss-Markov process with order less than \( K \), cf. Section 11.6 of Cover and Thomas [27]. The problem of describing all spectral densities \( S(e^{j\theta}) = \sum_l S_l e^{j\theta l} \) such that the Fourier coefficients \( S_l = S_{-l} \) take on specified values for \( 0 \leq l < K \) is essentially the Carathéodory-Toeplitz extension problem of complex function theory.

The capacity of a channel with additive colored gaussian noise when the input spectrum \( S_X(e^{j\theta}) \) is constrained to some convex set is attained by maximizing the mutual information rate between the input process and the output process or equivalently by maximizing the entropy rate of the output process. It suffices to consider gaussian input processes, because replacing any input by gaussian input with the same spectrum yields gaussian output with the same spectrum but higher entropy rate. The power spectral density of the output process after equalization is the sum of the input spectrum and the independent noise spectrum \( S_N(e^{j\theta}) \):

\[
S_Y(e^{j\theta}) = S_X(e^{j\theta}) + S_N(e^{j\theta}).
\]

The mutual information rate between the input and output processes is given by

\[
\lim_{n \to \infty} \frac{1}{n} I(X^n, Y^n) = \frac{1}{2} \int_{-\pi}^{\pi} \log \left( \frac{S_Y(e^{j\theta})}{S_N(e^{j\theta})} \right) d\theta.
\]

The optimal input spectrum \( \Sigma_X(e^{j\theta}) \) satisfies the KKT conditions: for any input spectrum \( S_X(e^{j\theta}) \) that satisfies the constraints, we have

\[
\int_{-\pi}^{\pi} \frac{S_X(e^{j\theta}) - \Sigma_X(e^{j\theta})}{S_N(e^{j\theta})} d\theta \leq 0.
\]

If the total input power \( \int_{-\pi}^{\pi} S_X(e^{j\theta}) d\theta/(2\pi) \) is not allowed to exceed some number \( P \), then the KKT conditions express the impossibility of transporting input power away from frequencies \( \theta \) where the output precision \((\Sigma_X(e^{j\theta}) + S_N(e^{j\theta}))^{-1}\) is low toward frequencies where this precision is high. Thus the output spectrum must take on a constant minimax value at frequencies where input power is present. The optimal input spectrum is obtained by water filling, as explained in Section 10.5 of Cover and Thomas [27]. In fact,

\[
\Sigma_X(e^{j\theta}) = [K - S_N(e^{j\theta})]^+,
\]

where the constant \( K \) is chosen so that \( \int \Sigma_X(e^{j\theta}) d\theta/(2\pi) = P \). The log-optimum input spectrum \( \Sigma_X(e^{j\theta}) \) is supported by \{ \theta : S_N(e^{j\theta}) \leq K \}, and the corresponding output spectrum \( \Sigma_X(e^{j\theta}) + S_N(e^{j\theta}) \) has a constant level \( K \) in this frequency band.

If \( G \) is a countable abelian group with the discrete topology, then the dual group \( \hat{G} \) of characters is compact and Haar measure on \( \hat{G} \) is a normalized probability measure \( P \). The Fourier transform establishes an isomorphism between the group algebra \( \ell^1(G) \) with convolution and the algebra \( C(\hat{G}) \) of continuous functions on \( \hat{G} \) with multiplication. By Bochner's theorem, an element of \( \ell^1(G) \) is positive definite iFF its spectrum is a nonnegative
function on $\mathcal{X} = \hat{G}$. One may wish to maximize $E_P\{\log q(X)\}$ over a convex set $Q$ of spectra $q(x) \geq 0$. More generally, if $\mathcal{A}$ is a commutative $C^*$-algebra then the space $\mathcal{X}$ of characters is a locally compact space, and the Gelfand transform establishes an isomorphism between $\mathcal{A}$ and the space of continuous functions on $\mathcal{X}$ that vanish at infinity. A probability measure $P(dx)$ on $\mathcal{X}$ specifies how to select a spectral component $X \in \mathcal{X}$ at random, and one may wish to maximize the geometric mean $\exp[E_P\{\log q(X)\}]$ over a convex set $Q$ of nonnegative spectra. If $\mathcal{A}$ is not commutative, then we get into the realm of quantum probability and quantum entropy theory. For this we refer to Ohya and Petz [97].

C. Maximizing the determinant of nonsymmetric matrices

Lewis [85] has studied the problem of maximizing $|\det S|$ subject to $\rho(S) \leq 1$ when $\rho(\cdot)$ is an arbitrary norm on the space of real $n \times n$ matrices. The matrices need not be symmetric. By compactness of the unit ball in $\mathbb{R}^{n \times n}$, there is a matrix $\Sigma$ with $\rho(\Sigma) \leq 1$ that maximizes $|\det S|$ subject to $\rho(S) \leq 1$. For any matrix $S$ and small $\epsilon$ we then have

$$\left|\det \left( \frac{\Sigma + \epsilon S}{\rho(\Sigma + \epsilon S)} \right) \right| \leq |\det \Sigma|,$$

and consequently, since $\rho(\Sigma + \epsilon S) \leq \rho(\Sigma) + \epsilon \rho(S)$ and $\rho(\Sigma) \leq 1$,

$$|\det(I + \epsilon \Sigma^{-1}S)| \leq \rho(\Sigma + \epsilon S)^n \leq (1 + \epsilon \rho(S))^n.$$

Taking the derivative of the left and right hand sides at $\epsilon = 0^+$, we see that

$$\text{tr} [\Sigma^{-1}S] \leq n \rho(S), \quad S \in \mathbb{R}^{n \times n}.$$

This means that $\rho^*(\Sigma^{-1}) \leq n$, where $\rho^*(\cdot)$ is the conjugate norm of $\rho(\cdot)$:

$$\rho^*(T) = \sup \{\text{tr}[TS] : S \in \mathbb{R}^{n \times n}, \rho(S) \leq 1\}.$$

Clearly $\text{tr}[TS] \leq \rho^*(T) \rho(S)$, and for $S = \Sigma$ and $T = \Sigma^{-1}$ we have

$$n = \text{tr}[\Sigma^{-1}\Sigma] \leq \rho^*(\Sigma^{-1}) \rho(\Sigma) \leq n.$$

We may conclude that $\rho(\Sigma) = 1$ and $\rho^*(\Sigma^{-1}) = n$. For further discussion see Chapter III of Pisier [102] and Chapter 6 of Diestel, Jarchow and Tonge [43].

As an illustration, we mention the problem of inscribing the ellipsoid with maximum volume inside the unit ball of an $n$-dimensional normed vector space $E$. Let $\| \cdot \|$ denote the norm on $E$, and let $\| \cdot \|_2$ denote the standard euclidean norm on $\mathbb{R}^n$. An ellipsoid in $E$ can be represented as the image of the unit ball $\{ x \in \mathbb{R}^n \mid \|x\|_2 \leq 1 \}$ via a linear mapping $S : \mathbb{R}^n \to E$. This ellipsoid is contained in the unit ball of $E$ iff $\|Sx\| \leq 1$ whenever $\|x\|_2 \leq 1$, or equivalently iff $\rho(S) \leq 1$ where $\rho(S)$ is the operator norm of $S$:

$$\rho(S) = \sup_{\|x\|_2 \leq 1} \|Sx\||.$$

The volume of the ellipsoid $\{Sx \mid \|x\|_2 \leq 1\}$ is proportional to $|\det S|$ since we may identify $E$ with $\mathbb{R}^n$ by choosing a basis in $E$. Thus maximizing the volume of the ellipsoid amounts to maximizing $|\det S|$ subject to $\rho(S) \leq 1$.

D. Convex corners defined by unitarily invariant matrix norms

78
The space $\mathcal{H}^n$ of complex hermitian $n \times n$ matrices is a real Hilbert space with inner product $(R, S)$ defined as $\text{tr}[RS]$. A function $F(S)$ on $\mathcal{H}^n$ is called unitarily invariant if $F(U^*SU) = F(S)$ for all $S \in \mathcal{H}^n$ and all unitary $n \times n$ matrices $U$. It is easy to see that $F(S)$ is unitarily invariant iff it is a function $f(\lambda(S))$ of the decreasing sequence $\lambda(S)$ of eigenvalues of the matrix $S$. Note that $F(\lambda) = F(\text{diag}(\lambda))$ where $\text{diag}(\lambda)$ is the diagonal matrix with diagonal entries $\lambda = (\lambda_1, \ldots, \lambda_n)$. A set $S$ of matrices in $\mathcal{H}^n$ is called unitarily invariant if $U^*SU \in S$ whenever $S \in S$ and $U$ is unitary. Such a set $S$ is closed and convex in $\mathcal{H}^n$ iff the set $\lambda(S) = \{\lambda(S) : S \in S\}$ is closed and convex in $\mathcal{R}^n$. See Lewis [83].

A spectral function $F(S) = f(\lambda(S))$ is convex on $\mathcal{H}^n$ iff $f(\lambda)$ is symmetric and convex on $\mathcal{R}^n$, or equivalently iff the restriction of $f$ to the cone $\{\lambda \in \mathcal{R}^n : \lambda_1 \geq \cdots \geq \lambda_n\}$ is both convex and Schur convex. An important class of symmetric convex functions are the symmetric gauge functions. These are precisely the functions $f(\lambda)$ on $\mathcal{R}^n$ such that $F(S) = f(\lambda(S))$ is a unitarily invariant norm on $\mathcal{H}^n$.

Let $Q_F = \{S \in \mathcal{H}^n : S \geq 0, F(S) \leq 1\}$ be the set of positive semidefinite matrices in the unit ball for a unitarily invariant matrix norm $F(S) = f(\lambda(S))$ on $\mathcal{H}^n$. Then the image $\lambda(Q_F) = \{\lambda(S) : S \in Q_F\}$ is equal to the convex corner $Q_f = \{\lambda \in \mathcal{R}^+_n : f(\lambda) \leq 1\}$ defined by the symmetric gauge function $f(\lambda)$. The polar of the convex corner $Q_f$ is equal to the convex corner $Q_g$ defined by the symmetric gauge function $g$ that is the polar of $f$. In terms of the inner product $(\mu, \lambda) = \sum_{1 \leq i \leq n} \mu_i \lambda_i$ on $\mathcal{R}^n$, the polar $g(\mu)$ is defined as

$$g(\mu) = \sup_{\lambda \in Q_f} (\mu, \lambda).$$

The polar convex corner $Q_g$ is equal to the image $\lambda(Q_G)$ of the positive semidefinite part $Q_G$ of the unit ball for the unitarily invariant norm $G(R)$ which is the polar of $F(S)$:

$$G(R) = \sup_{S \in Q_F} (R, S).$$

This follows from the inequality $(R, S) \leq (\lambda(R), \lambda(S))$, which is related to a result of von Neumann [125] that we shall state below. Theobald [117] proved that the inequality holds with equality iff $R$ and $S$ can be simultaneously diagonalized by a unitary matrix $U$ such that $U^*RU = \text{diag}(\lambda(R))$ and $U^*SU = \text{diag}(\lambda(S))$.

One may wish to maximize the concave function $\log \det S$ over a compact convex unitarily invariant set $S$ of positive semidefinite matrices in $\mathcal{H}^n$. This amounts to maximizing the volume $\prod_{1 \leq i \leq n} \lambda_i$ or the average $E_k \{\log \lambda_k\} = n^{-1} \sum_{1 \leq i \leq n} \log \lambda_i$ over the compact convex set $\lambda(S)$ in $\mathcal{R}^n$. This is a convex programming problem that often can be solved efficiently by interior point methods, cf. Vandenberghe, Boyd and Wu [120].

Besides $\Phi_0(S) = (\det(S))^{1/n}$, there are other concave unitarily invariant functions $\Phi(S) = \varphi(\lambda(S))$ that one may want to maximize over a unitarily invariant set $S$ of complex hermitian $n \times n$ matrices. Also, there are convex unitarily invariant functions that one may want to minimize. In particular, consider the matrix means

$$\Phi_p(S) = \varphi_p(\lambda(S)) = \left( \frac{1}{n} \sum_{1 \leq i \leq n} |\lambda_i(S)|^p \right)^{1/p}.$$

If $S$ is positive semidefinite then $\Phi_{\infty}(S) = \lambda_{\text{max}}(S)$ is the maximum, $\Phi_1(S) = n^{-1} \text{tr}[S]$ is the arithmetic mean, $\Phi_0(S) = (\det S)^{1/n}$ is the geometric mean, $\Phi_{-1}(S)$ is the harmonic
mean and $\Phi_{-\infty}(S) = \lambda_{\min}(S)$ is the minimum of the eigenvalues of $S$. If $p \geq 1$ then $\varphi_p(\lambda)$ and $\Phi_p(S)$ are scaled versions of the usual $p$-norm on $\mathcal{R}^n$ and the Schatten $p$-norm on $\mathcal{H}^n$. If $p \leq 1$ then $\Phi_p(\lambda)$ is concave (in fact, log-concave) in $\lambda$ and $\Phi_p(S)$ is sometimes called an information function. Such information functions are often maximized over a compact convex set of positive semidefinite information matrices (inverse covariance matrices) in the design of experiments, cf. Pukelsheim [104]. The quantities $\Phi_p(S) = \varphi_p(\lambda)$ are related to the means studied by Hardy, Littlewood and Pólya [64]. Also, $\log \Phi_p(S)$ is similar to the Rényi entropy of $\lambda(S)$. An interesting survey of eigenvalue optimization is given by Lewis and Overton [84].

All the above considerations apply to orthogonally invariant sets and functions of real symmetric $n \times n$ matrices. Also, an analogous theory can be developed for the real Hilbert space of all complex $m \times n$ matrices with inner product $(R, S) = \Re \text{tr}[RS^*]$. A function $F(S)$ is now called unitarily invariant if $F(USV) = F(S)$ for all $S$ and all unitary $m \times m$ matrices $U$ and unitary $n \times n$ matrices $V$. Such functions can be written as $F(S) = f(\sigma(S))$ where $\sigma(S)$ is the descending sequence of singular values of $S$ and $f(\sigma)$ is a symmetric and absolute function of $\sigma$ (absolute means that $f(\sigma) = f(|\sigma_1|, \ldots, |\sigma_n|)$ when $\sigma = (\sigma_1, \ldots, \sigma_n)$). Maximizing $|\log \det S|$ over a unitarily invariant set $S$ amounts to maximizing $\prod_i \sigma_i$ over the set $\sigma(S) = \{\sigma \in \mathcal{R}^n_+ : \sigma = \sigma(S) \text{ for some } S \in \mathcal{S}\}$.

The function $F(S) = f(\sigma(S))$ is a unitarily invariant matrix norm iff $f(\sigma)$ is a symmetric gauge function on $\mathcal{R}^{m\times n}$. This is a result of von Neumann [125], who proved that $(R, S) \leq (\sigma(R), \sigma(S))$ when $R$ and $S$ are $n \times n$ matrices. The inequality also holds for $m \times n$ matrices, and it holds with equality iff there are unitary matrices $U$ and $V$ such that $URV = \text{diag} \sigma(R)$ and $USV = \text{diag} \sigma(S)$. It follows that the convex corner $Q_f$ and its polar $Q_g$ are the image $Q_f = \sigma(Q_F)$ and $Q_g = \sigma(Q_G)$ of the positive semidefinite part of the unit ball for the unitarily invariant matrix norm $F(S) = f(\sigma(S))$ and its polar $G(R) = g(\sigma(R))$. For a study of the faces and exposed faces of the convex corners $Q_f$, $Q_g$, $Q_F$ and $Q_G$ see de Sá [40] [41] and the literature that is cited there.

The theory can be generalized to bounded linear operators on infinite dimensional Hilbert spaces. By constraining the sequence of s-numbers of operators to a symmetric sequence space in the sense of Köthe and Toeplitz [75], one obtains an operator ideal with a unitarily invariant norm. Such symmetrically-normed operator ideals were studied by von Neumann and Schatten [112], Gohberg and Krein [57], Pietsch [99], Simon [114], and others. For interpolation of operator ideals see Pietsch and Triebel [100] and Arazy [5]. Uhlmann [119] and Kosaki [74] develop an interpolation theory for seminorms to define the relative entropy between two positive semidefinite operators. Also relevant is work on interpolation of operators on Banach lattices and rearrangement invariant function spaces, cf. Bennett and Sharpley [13] and Brudnyi, Krein and Semenov [21].

E. Products of random matrices

Let $S$ be a random real $d \times d$ matrix with positive determinant, and let $\{S_i\}$ be a sequence of independent realizations of $S$. If the matrix transformations $S_0, \ldots, S_{n-1}$ are sequentially applied to row vectors in the unit ball $\{\xi \in \mathcal{R}^d : ||\xi|| \leq 1\}$, then the volume of the resulting ellipsoid is proportional to the determinant of the product $S^n = S_0 S_1 \ldots S_{n-1}$. Clearly, $\det S^n$ grows exponentially fast almost surely with limiting rate $E[\log \det S]$. In fact, the multiplicative ergodic theorem for products of random matrices (e.g. see Ruelle [109]) implies the existence of a symmetric positive semidefinite $d \times d$ matrix $\Gamma$ such that

$$[S^n S^*]^{\frac{1}{n}} \to \Gamma \quad \text{almost surely.}$$
The logarithms of the eigenvalues of the matrix $\Gamma$ are called the Lyapunov or characteristic exponents of the random matrix $S$. If all Lyapunov exponents are finite then

$$\lim_{n \to \infty} \frac{1}{n} \log \det S^n = E\{\log \det S\} = \log \det \Gamma \quad \text{almost surely.}$$

Thus $\det S^n$ grows exponentially with limiting rate $\log \det \Gamma = \text{tr} \log \Gamma$, that is the sum of the Lyapunov exponents counted according to their multiplicity.

One may wish to maximize the exponential growth rate $E\{\log \det S\}$ over a convex family $S$ of random $d \times d$ matrices. Suppose $\Sigma$ attains the maximum growth exponent

$$E\{\log \det \Sigma\} = \sup_{S \in S} E\{\log \det S\}.$$

The function $\log \det S$ is concave in $S$, hence $\Sigma$ is characterized by the KKT conditions $D(S|\Sigma) \leq 0$ for all $S \in S$. Here $D(S_1|S_0)$ denotes the directional derivative:

$$D(S_1|S_0) = \left. \frac{d}{d\epsilon} E\{\log \det S_\epsilon\} \right|_{\epsilon=0+},$$

where $S_\epsilon = (1-\epsilon)S_0 + \epsilon S_1$. Recall that when $S_0$ is non-singular, $\frac{d}{d\epsilon} \log \det S_\epsilon = \text{tr}[(S_1 - S_0)S_0^{-1}] = \text{tr}[S_1S_0^{-1}] - d$,

where $d = \text{tr}[S_0S_0^{-1}]$ is the trace of the $d \times d$ identity matrix. It follows that

$$D(S_1|S_0) = \text{tr} E\{S_1S_0^{-1}\} - d.$$

The KKT conditions amount to $\text{tr} E\{(S - \Sigma)\Sigma^{-1}\} \leq 0$ or

$$\text{tr} E\{S\Sigma^{-1}\} \leq d, \quad \text{for all } S \in S.$$
References


