

Article

Divergence and Sufficiency for Convex Optimization

Peter Harremoës

GSK Department, Copenhagen Business College, Nørre Voldgade 34, 1358 Copenhagen K, Denmark; harremoes@iee.org; Tel.: +45-39-56-41-71

Academic Editor: Renaldas Urniezius

Received: 30 December 2016; Accepted: 2 May 2017; Published: 3 May 2017

Abstract: Logarithmic score and information divergence appear in information theory, statistics, statistical mechanics, and portfolio theory. We demonstrate that all these topics involve some kind of optimization that leads directly to regret functions and such regret functions are often given by Bregman divergences. If a regret function also fulfills a sufficiency condition it must be proportional to information divergence. We will demonstrate that sufficiency is equivalent to the apparently weaker notion of locality and it is also equivalent to the apparently stronger notion of monotonicity. These sufficiency conditions have quite different relevance in the different areas of application, and often they are not fulfilled. Therefore sufficiency conditions can be used to explain when results from one area can be transferred directly to another and when one will experience differences.

Keywords: Bregman divergence; entropy; exergy; Kraft's inequality; locality; monotonicity; portfolio; regret; scoring rule; sufficiency

1. Introduction

One of the main purposes of information theory is to compress data so that data can be recovered exactly or approximately. One of the most important quantities was called entropy because it is calculated according to a formula that mimics the calculation of entropy in statistical mechanics. Another key concept in information theory is information divergence (KL-divergence) that is defined for probability vectors P and Q as

$$D(P\|Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}.$$

It was introduced by Kullback and Leibler in 1951 in a paper entitled On Information and Sufficiency [1]. The link from information theory back to statistical physics was developed by E.T. Jaynes via the maximum entropy principle [2–4]. The link back to statistics is now well established [5–9].

Related quantities appear in information theory, statistics, statistical mechanics, and finance, and we are interested in a theory that describes when these relations are exact and when they just work by analogy. First we introduce some general results about optimization on state spaces of finite dimensional C^* -algebras. This part applies exactly to all the topics under consideration and lead to Bregman divergences or more general regret functions. Secondly, we introduce several notions of sufficiency and show that this leads to information divergence. In a number of cases it is not possible or not relevant to impose the condition of sufficiency, which can explain why regret function are not always equal to information divergence.

2. Structure of the State Space

Our knowledge about a system will be represented by a state space. In many applications the state space is given by a set of probability distributions on a sample space. In such cases the state

space is a simplex, but it is well-known that the state space is not a simplex in quantum physics. For applications in quantum physics the state space is often represented by a set of density matrices, i.e., positive semidefinite complex matrices with trace 1. In some cases the states are represented as elements of a finite dimensional C^* -algebra, which is a direct sum of matrix algebras. A finite dimensional C^* -algebra that is a sum of 1×1 matrices has a state space that is a simplex, so the state spaces of finite dimensional C^* -algebras contain the classical probability distributions as special cases.

The extreme points in the set of states are the pure states. The pure states of a C^* -algebra can be identified with projections of rank 1. Two density matrices s_1 and s_2 are said to be orthogonal if $s_1 s_2 = s_2 s_1 = 0$. Any state s has a decomposition

$$s = \sum \lambda_i s_i$$

where s_i are orthogonal pure states. Such a decomposition is not unique, but for a finite dimensional C^* -algebra the coefficients $\lambda_1, \lambda_2, \dots, \lambda_n$ are unique and are called the spectrum of the state.

Sometimes more general state spaces are of interest. In generalized probabilistic theories a state space is a convex set where mixtures are defined by randomly choosing certain states with certain probabilities [10,11]. A convex set where all orthogonal decompositions of a state have the same spectrum, is called a spectral state space. Much of the theory in this paper can be generalized to spectral sets. The most important spectral sets are sets of positive elements with trace 1 in Jordan algebras. The study of Jordan algebras and other spectral sets is relevant for the foundation of quantum theory [12–15], but in this paper we will restrict our attention to states on finite dimensional C^* -algebras. Nevertheless some of the theorems and proofs are stated in such a way that they hold for more general state spaces.

3. Optimization

Let \mathcal{S} denote a state space of a finite dimensional C^* -algebra and let \mathcal{A} denote a set of self-adjoint operators. Each $a \in \mathcal{A}$ is identified with a real valued measurement. The elements of \mathcal{A} may represent feasible *actions* (decisions) that lead to a payoff like the score of a statistical decision, the energy extracted by a certain interaction with the system, (minus) the length of a codeword of the next encoded input letter using a specific code book, or the revenue of using a certain portfolio. For each $s \in \mathcal{S}$ the mean value of the measurement $a \in \mathcal{A}$ is given by

$$\langle a, s \rangle = \text{tr}(as).$$

In this way the set of actions may be identified with a subset of the dual space of \mathcal{S} .

Next we define

$$F(s) = \sup_{a \in \mathcal{A}} \langle a, s \rangle.$$

We note that F is convex, but F need not be strictly convex. In principle $F(s)$ may be infinite, but we will assume that $F(s) < \infty$ for all states s . We also note that F is lower semi-continuous. In this paper we will assume that the function F is continuous. The assumption that F is a real valued continuous function is fulfilled for all the applications we consider.

If s is a state and $a \in \mathcal{A}$ is an action then we say that a is *optimal* for s if $\langle a, s \rangle = F(s)$. A sequence of actions $a_n \in \mathcal{A}$ is said to be *asymptotically optimal* for the state s if $\langle a_n, s \rangle \rightarrow F(s)$ for $n \rightarrow \infty$.

If a_i are actions and (t_i) is a probability vector then we may define the mixed action $\sum t_i \cdot a_i$ as the action where we do the action a_i with probability t_i . We note that $\langle \sum t_i \cdot a_i, s \rangle = \sum t_i \cdot \langle a_i, s \rangle$. We will assume that all such mixtures of feasible actions are also feasible. If $a_1(s) \geq a_2(s)$ almost surely for all states we say that a_1 *dominates* a_2 and if $a_1(s) > a_2(s)$ almost surely for all states s we say that a_1 *strictly dominates* a_2 . All actions that are dominated may be removed from \mathcal{A} without changing the function F . Let \mathcal{A}_F denote the set of self-adjoint operators (observables) a such that

$\langle a, s \rangle \leq F(s)$. Then $F(s) = \sup_{a \in \mathcal{A}_F} \langle a, s \rangle$. Therefore we may replace \mathcal{A} by \mathcal{A}_F without changing the optimization problem.

In the definition of regret we follow Servage [16] but with different notation.

Definition 1. Let F denote a convex function on the state space \mathcal{S} . If $F(s)$ is finite the regret of the action a is defined by

$$D_F(s, a) = F(s) - \langle a, s \rangle. \tag{1}$$

The notion of regret has been discussed in detail in [17–19]. In [20] it was proved that if a regret based decision procedure is transitive then it must be equal to a difference in expected utility as in Equation (1), which rules out certain non-linear models in [17,19].

Proposition 1. The regret D_F of actions has the following properties:

- $D_F(s, a) \geq 0$ with equality if a is optimal for s .
- $s \rightarrow D_F(s, a)$ is a convex function.
- If \bar{a} is optimal for the state $\bar{s} = \sum t_i \cdot s_i$ where $(t_1, t_2, \dots, t_\ell)$ is a probability vector then

$$\sum t_i \cdot D_F(s_i, a) = \sum t_i \cdot D_F(s_i, \bar{a}) + D_F(\bar{s}, a).$$

- $\sum t_i \cdot D_F(s_i, a)$ is minimal if a is optimal for $\bar{s} = \sum t_i \cdot s_i$.

If the state is s_1 but one acts as if the state were s_0 one may compare what one achieves and what could have been achieved. If the state s_0 has a unique optimal action a we may simply define the regret of s_0 by

$$D_F(s_1, s_0) = D_F(s_1, a).$$

The following definition leads to a regret function that is essentially equivalent to the so-called generalized Bregman divergences defined by Kiwiel [21,22].

Definition 2. Let F denote a convex function on the state space \mathcal{S} . If $F(s_1)$ is finite then we define the regret of the state s_0 as

$$D_F(s_1, s_0) = \inf_{(a_n)} \lim_{n \rightarrow \infty} D_F(s_1, a_n)$$

where the infimum is taken over all sequences of actions (a_n) that are asymptotically optimal for s_0 .

With this definition the regret is always defined with values in $[0, \infty]$ and the value of the regret $D_F(s_1, s_0)$ only depends on the restriction of the function F to the line segment from s_0 to s_1 . Let f denote the function $f(t) = F((1-t)s_0 + ts_1)$ where $t \in [0, 1]$. As illustrated in Figure 1 we have

$$D_F(s_1, s_0) = f(1) - (f(0) + f'_+(0)) \tag{2}$$

where $f'_+(0)$ denotes the right derivative of f at $t = 0$. Equation (2) is even valid when the regret is infinite if we allow the right derivative to take the value $-\infty$.

If the state s_0 has the unique optimal action $a \in \mathcal{A}$ then

$$F(s_1) = D_F(s_1, s_0) + \langle a, s_1 \rangle \tag{3}$$

so the function F can be reconstructed from D_F except for an affine function of s_1 . The following proposition follows from Alexandrov’s theorem ([23], Theorem 25.5).

Proposition 2. A convex function on a finite dimensional convex set is differentiable almost everywhere with respect to the Lebesgue measure.

A state s_0 where F is differentiable has a unique optimal action. Therefore Equation (3) holds for almost any state s_0 . In particular the function F can be reconstructed from D_F except for an affine function.

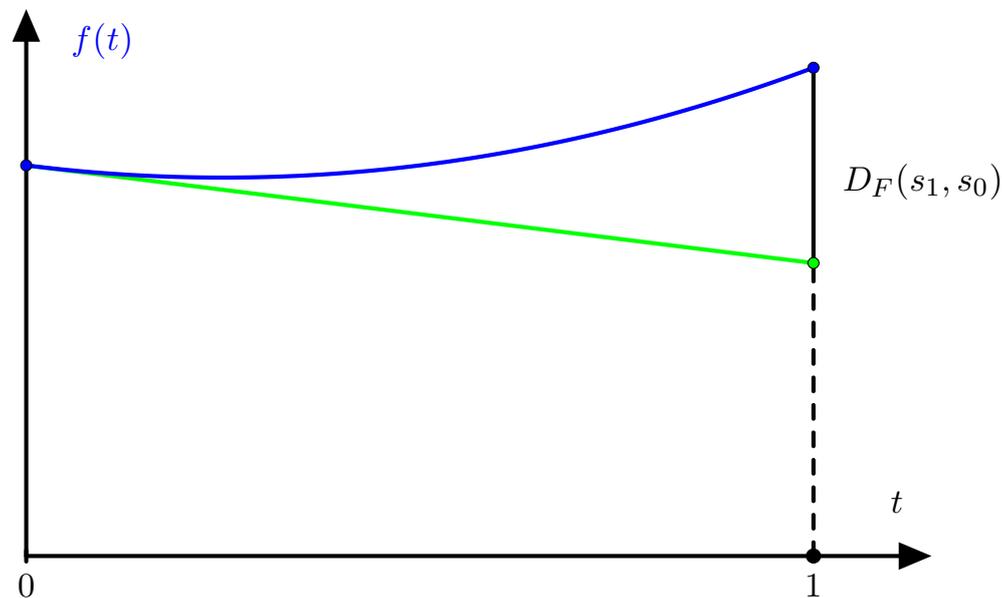


Figure 1. The regret equals the vertical distance between curve and tangent.

Proposition 3. The regret D_F of states has the following properties:

- $D_F(s_1, s_0) \geq 0$ with equality if there exists an action a that is optimal for both s_1 and s_0 .
- $s_1 \rightarrow D_F(s_1, s_0)$ is a convex function.

Further, the following two conditions are equivalent:

- $D_F(s_1, s_0) = 0$ implies $s_1 = s_0$.
- The function F is strictly convex.

We say that a regret function D_F is strict if F is strictly convex. The two last properties Proposition 1 do not carry over to regret for states except if the regret is a Bregman divergence as defined below. The regret is called a Bregman divergence if it can be written in the following form

$$D_F(s_1, s_0) = F(s_1) - (F(s_0) + \langle s_1 - s_0, \nabla F(s_0) \rangle) \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the (Hilbert-Schmidt) inner product. In the context of forecasting and statistical scoring rules the use of Bregman divergences dates back to [24]. A similar but less general definition of regret was given by Rao and Nayak [25] where the name *cross entropy* was proposed. Although Bregman divergences have been known for many years they did not gain popularity before the paper [26] where a systematic study of Bregman divergences was presented.

We note that if D_F is a Bregman divergence and s_0 minimizes F then $\nabla F(s_0) = 0$ so that the formula for the Bregman divergence reduces to

$$D_F(s_1, s_0) = F(s_1) - F(s_0).$$

Bregman divergences satisfy the Bregman identity

$$\sum t_i \cdot D_F(s_i, s) = \sum t_i \cdot D_F(s_i, \bar{s}) + D_F(\bar{s}, s), \tag{5}$$

but if F is not differentiable this identity can be violated.

Example 1. Let the state space be the interval $[0, 1]$ with two actions $\langle a_0, s \rangle = 1 - 2s$ and $\langle a_1, s \rangle = 2s - 1$. Let $s_0 = 0$ and $s_1 = 1$. Let further $t_0 = 1/3$ and $t_1 = 2/3$. Then $\bar{s} = 2/3$. If $s = 1/2$ then

$$\sum t_i \cdot D_F(s_i, s) = 0,$$

but

$$\begin{aligned} \sum t_i \cdot D_F(s_i, \bar{s}) &= \frac{1}{3} \cdot (\langle a_0, 0 \rangle - \langle a_1, 0 \rangle) + \frac{2}{3} \cdot (\langle a_1, 1 \rangle - \langle a_1, 1 \rangle) \\ &= \frac{1}{3} \cdot (1 - (-1)) \\ &= \frac{2}{3}. \end{aligned}$$

Clearly the Bregman identity (5) is violated and $\sum t_i \cdot D_F(s_i, s)$ will increase if s is replaced by \bar{s} .

The following proposition is easily proved.

Proposition 4. For a convex and continuous function F on the state space S the following conditions are equivalent:

- The function F is differentiable in the interior of any face of S .
- The regret D_F is a Bregman divergence.
- The Bregman identity (5) is always satisfied.
- For any probability vectors (t_1, t_2, \dots, t_n) the sum $\sum t_i \cdot D_F(s_i, s)$ is always minimal when $s = \sum t_i \cdot s_i$.

4. Examples

In this section we shall see how regret functions are defined in some applications.

4.1. Information Theory

We recall that a code is uniquely decodable if any finite sequence of input symbols give a unique sequence of output symbols. It is well-known that a uniquely decodable code satisfies Kraft's inequality (see [27] and ([28], Theorem 3.8))

$$\sum_{a \in \mathbb{A}} \beta^{-\ell(a)} \leq 1 \tag{6}$$

where $\ell(a)$ denotes the length of the codeword corresponding to the input symbol $a \in \mathbb{A}$ and β denotes the size of the output alphabet \mathbb{B} . Here the length of a codeword is an integer. If $P = (p_a)_{a \in \mathbb{A}}$ is a probability vector over the input alphabet, then the mean code-length is

$$\sum_{a \in \mathbb{A}} \ell(a) \cdot p_a.$$

Our goal is to minimize the expected code-length. Here the state space consist of probability distributions over the input alphabet and the actions are code-length functions.

Shannon established the inequality

$$-\sum_{a \in \mathbb{A}} \log_b(p_a) \cdot p_a \leq \min_{a \in \mathbb{A}} \sum_{a \in \mathbb{A}} \ell(a) \cdot p_a \leq -\sum_{a \in \mathbb{A}} \log_b(p_a) \cdot p_a + 1.$$

It is a combinatoric problem to find the optimal code-length function. In the simplest case with a binary output alphabet the optimal code-length function is determined by the Huffmann algorithm.

A code-length function dominates another code-length function if all letters have shorter code-length. If a code-length function is not dominated by another code-length function then for all $a \in \mathbb{A}$ the length is bounded by $\ell(a) \leq |\mathbb{A}| - 1$. For fixed alphabets \mathbb{A} and \mathbb{B} there exists only a finite

number of code-length functions ℓ that satisfy Kraft's inequality and are not dominated by other code-length functions that satisfying Kraft's inequality.

4.2. Scoring Rules

The use of scoring rules has a long history in statistics. An early contribution was the idea of minimizing the sum of square deviations that dates back to Gauss and works perfectly for Gaussian distributions. In the 1920s, Ramsay and de Finetti proved versions of the Dutch book theorem where determination of probability distributions were considered as dual problems of maximizing a payoff function [29]. Later it was proved that any consistent inference procedure corresponds to optimizing with respect to some payoff function. A more systematic study of scoring rules was given by McCarthy [30].

Consider an experiment with $\mathcal{X} = \{1, 2, \dots, \ell\}$ as sample space. A *scoring rule* f is defined as a function $\mathcal{X} \times M_1^+(\mathcal{X}) \rightarrow \mathbb{R}$ such that the score is $f(x, Q)$ when a prediction has been given in terms of a probability distribution Q and $x \in \mathcal{X}$ has been observed. A scoring rule is *proper* if for any probability measure $P \in M_1^+(\mathcal{X})$ the score $\sum_{x \in \mathcal{X}} P(x) \cdot f(x, Q)$ is minimal when $Q = P$. Here the state space consist of probability distributions over \mathcal{X} and the actions are predictions over \mathcal{X} , which are also probability distributions over \mathcal{X} .

There is a correspondence between proper scoring rules and Bregman divergences as explained in [31,32]. If D_F is a Bregman divergence and g is a function with domain \mathcal{X} then f given by $f(x, Q) = g(x) - D_F(\delta_x, Q)$ defines a scoring rule.

Assume that f is a proper scoring function. Then a function F can be defined as

$$F(P) = \sum_{x \in \mathcal{X}} P(x) \cdot f(x, P).$$

This lead to the regret function

$$D_F(P, Q) = F(P) - \sum_{x \in \mathcal{X}} P(x) \cdot f(x, Q). \quad (7)$$

Since f is assumed to be proper $D_F(P, Q) \geq 0$. The Bregman identity (5) follows by straight forward calculations. With these two results we see that the regret function D_F is a Bregman divergence and that

$$\begin{aligned} D_F(\delta_y, Q) &= \sum_{x \in \mathcal{X}} \delta_y(x) \cdot f(x, \delta_y) - \sum_{x \in \mathcal{X}} \delta_y(x) \cdot f(x, Q) \\ &= f(y, \delta_y) - f(y, Q). \end{aligned} \quad (8)$$

Hence a proper scoring rule f has the form $f(x, Q) = g(x) - D_F(\delta_x, Q)$ where $g(x) = f(x, \delta_x)$. A *strictly proper scoring rule* can be defined as a proper scoring rule where the corresponding Bregman divergence is strict.

Example 2. The Brier score is given by

$$f(x, Q) = \frac{1}{n} \left(\sum_{y \in \mathcal{X}} (Q(y) - \delta_x(y))^2 \right).$$

The Brier score is generated by the strictly convex function $F(P) = \frac{1}{n} \sum_{x \in \mathcal{X}} P(x)^2$.

4.3. Statistical Mechanics

Thermodynamics is the study of concepts like heat, temperature and energy. A major objective is to extract as much energy from a system as possible. The idea in statistical mechanics is to view the macroscopic behavior of a thermodynamic system as a statistical consequence of the interaction between a lot of microscopic components where the interacting between the components are governed by very simple laws. Here the central limit theorem and large deviation theory play a major role. One of the main achievements is the formula for entropy as a logarithm of a probability.

Here we shall restrict the discussion to the most simple kind of thermodynamic system from which we want to extract energy. We may think of a system of non-interacting spin particles in a magnetic field. For such a system the Hamiltonian is given by

$$\hat{H}(\sigma) = -\mu \sum h_j \sigma_j$$

where σ is the spin configuration, μ is the magnetic moment, h_j is the strength of an external magnetic field, and $\sigma_j = \pm 1$ is the spin of the the j 'th particle. If the system is in thermodynamic equilibrium the configuration probability is

$$P_\beta(\sigma) = \frac{\exp(-\beta \hat{H}(\sigma))}{Z_\beta}$$

where $Z(\beta)$ is the partition function

$$Z(\beta) = \sum_{\sigma} \exp(-\beta \hat{H}(\sigma)).$$

Here β is the inverse temperature $(kT)^{-1}$ of the spin system and $k = 1.381 \times 10^{-23}$ J/K is Boltzmann's constant.

The mean energy is given by

$$\sum_{\sigma} P_\beta(\sigma) \hat{H}(\sigma),$$

which will be identified with the internal energy U defined in thermodynamics. The Shannon entropy can be calculated as

$$\begin{aligned} -\sum_{\sigma} P_\beta(\sigma) \ln P_\beta(\sigma) &= -\sum_{\sigma} P_\beta(\sigma) \ln \frac{\exp(-\beta \hat{H}(\sigma))}{Z_\beta} \\ &= -\sum_{\sigma} P_\beta(\sigma) (-\beta \hat{H}(\sigma) - \ln Z(\beta)) \\ &= \beta \cdot U + \ln Z(\beta). \end{aligned}$$

The Shannon entropy times k will be identified with the thermodynamic entropy S .

The amount of energy that can be extracted from the system if a heat bath is available, is called the *exergy* [33]. We assume that the heat bath has temperature T_0 and the internal energy and entropy of the system are U_0 and S_0 if the system has been brought in equilibrium with the heat bath. The exergy can be calculated by

$$\begin{aligned} Ex &= U - U_0 - T_0(S - S_0) \\ &= U - U_0 - kT_0(\beta \cdot U + \ln Z(\beta) - \beta_0 U_0 - \ln Z(\beta_0)) \\ &= kT_0 \left((\beta_0 - \beta) \cdot U + \ln \frac{Z(\beta_0)}{Z(\beta)} \right). \end{aligned}$$

The information divergence between the actual state and the corresponding state that is in equilibrium with the environment is

$$\begin{aligned}
 D(P_\beta \| P_{\beta_0}) &= \sum_{\sigma} P_\beta(\sigma) \ln \frac{P_\beta(\sigma)}{P_{\beta_0}(\sigma)} \\
 &= \sum_{\sigma} P_\beta(\sigma) \ln \frac{\frac{\exp(-\beta \hat{H}(\sigma))}{Z(\beta)}}{\frac{\exp(-\beta_0 \hat{H}(\sigma))}{Z(\beta_0)}} \\
 &= \sum_{\sigma} P_\beta(\sigma) \left(-\beta \hat{H}(\sigma) + \beta_0 \hat{H}(\sigma) + \ln \frac{Z(\beta_0)}{Z(\beta)} \right) \\
 &= (\beta_0 - \beta) \cdot \sum_{\sigma} P_\beta(\sigma) \hat{H}(\sigma) + \ln \frac{Z(\beta_0)}{Z(\beta)} \\
 &= (\beta_0 - \beta) \cdot U + \ln \frac{Z(\beta_0)}{Z(\beta)}.
 \end{aligned}$$

Hence

$$Ex = kT_0 D(P_\beta \| P_{\beta_0}).$$

This equation appeared already in [34].

4.4. Portfolio Theory

The relation between information theory and gambling was established by J. L. Kelly [35]. Logarithmic terms appear because we are interested in the exponent in the exponential growth rate of our wealth. Later Kelly’s approach has been generalized to trading of stocks although the relation to information theory is weaker [36].

Let X_1, X_2, \dots, X_k denote price relatives for a list of k assets. For instance $X_5 = 1.04$ means that 5-th asset increases its value by 4%. Such price relatives are mapped into a price relative vector $\vec{X} = (X_1, X_2, \dots, X_k)$.

Example 3. A special asset is the safe asset where the price relative is 1 for any possible price relative vector. Investing in this asset corresponds to placing the money at a safe place with interest rate equal to 0%.

A portfolio is a probability vector $\vec{b} = (b_1, b_2, \dots, b_k)$ where for instance $b_5 = 0.3$ means that 30% of the money is invested in asset no. 5. We note that a portfolio may be traded just like the original assets. The price relative for the portfolio \vec{b} is $X_1 \cdot b_1 + X_2 \cdot b_2 + \dots + X_k \cdot b_k = \langle \vec{X}, \vec{b} \rangle$. The original assets may be considered as extreme points in the set of portfolios. If an asset has the property that the price relative is only positive for one of the possible price relative vectors, then we may call it a gambling asset.

We now consider a situation where the assets are traded once every day. For a sequence of price relative vectors $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$ and a constant re-balancing portfolio \vec{b} the wealth after n days is

$$S_n = \prod_{i=1}^n \langle \vec{X}_i, \vec{b} \rangle \tag{9}$$

$$= \exp \left(\sum_{i=1}^n \ln \left(\langle \vec{X}_i, \vec{b} \rangle \right) \right) \tag{10}$$

$$= \exp \left(n \cdot E \left[\ln \langle \vec{X}, \vec{b} \rangle \right] \right) \tag{11}$$

where the expectation is taken with respect to the empirical distribution of the price relative vectors. Here $E \left[\ln \langle \vec{X}, \vec{b} \rangle \right]$ is proportional to the doubling rate and is denoted $W(\vec{b}, P)$ where P indicates the

probability distribution of \vec{X} . Our goal is to maximize $W(\vec{b}, P)$ by choosing an appropriate portfolio \vec{b} . The advantage of using constant rebalancing portfolios was demonstrated in [37].

Definition 3. Let \vec{b}_1 and \vec{b}_2 denote two portfolios. We say that \vec{b}_1 dominates \vec{b}_2 if $\langle \vec{X}_j, \vec{b}_1 \rangle \geq \langle \vec{X}_j, \vec{b}_2 \rangle$ for any possible price relative vector $\vec{X}_j, j = 1, 2, \dots, n$. We say that \vec{b}_1 strictly dominates \vec{b}_2 if $\langle \vec{X}_j, \vec{b}_1 \rangle > \langle \vec{X}_j, \vec{b}_2 \rangle$ for any possible price relative vector $\vec{X}_j, j = 1, 2, \dots, n$. A set A of assets is said to dominate the set of assets B if any asset in B is dominated by a portfolio of assets in A .

The maximal doubling rate does not change if dominated assets are removed. Sometimes assets that are dominated but not strictly dominated may lead to non-uniqueness of the optimal portfolio.

Let \vec{b}_P denote a portfolio that is optimal for P and define

$$G(P) = W(\vec{b}_P, P). \tag{12}$$

The regret of choosing a portfolio that is optimal for Q when the distribution is P is given by the regret function

$$D_G(P, Q) = W(\vec{b}_P, P) - W(\vec{b}_Q, P). \tag{13}$$

If \vec{b}_Q is not uniquely determined we take a minimum over all \vec{b} that are optimal for Q .

Example 4. Assume that the price relative vector is $(2, 1/2)$ with probability $1 - t$ and $(1/2, 2)$ with probability t . Then the portfolio concentrated on the first asset is optimal for $t \leq 1/5$ and the portfolio concentrated on the second asset is optimal for $t > 4/5$. For values of t between $1/5$ and $4/5$ the optimal portfolio invests money on both assets as illustrated in Figure 2.

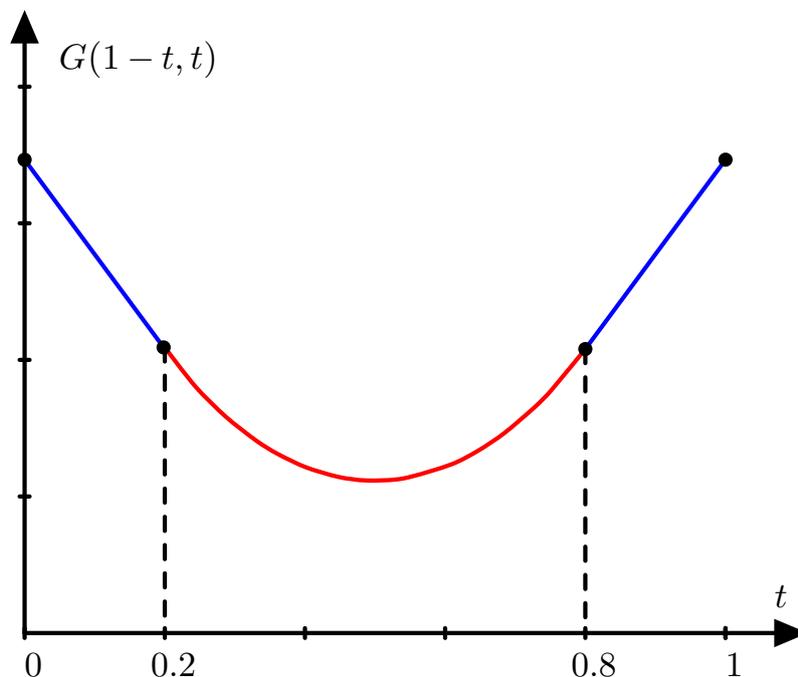


Figure 2. The function G for the price relative vectors in Example 4.

Lemma 1. If there are only two price relative vectors and the regret function is strict then either one of the assets dominates all other assets or two of the assets are orthogonal gambling assets that dominate all other assets.

Proof. We will assume that no assets are dominated by other assets. Let

$$\begin{aligned} \vec{X} &= (X_1, X_2, \dots, X_k) \\ \vec{Y} &= (Y_1, Y_2, \dots, Y_k) \end{aligned}$$

denote the two price relative vectors. Without loss of generality we may assume that

$$\frac{X_1}{Y_1} \geq \frac{X_2}{Y_2} \geq \dots \geq \frac{X_k}{Y_k}.$$

If $\frac{X_i}{Y_i} = \frac{X_{i+1}}{Y_{i+1}}$ then $\frac{X_i}{X_{i+1}} = \frac{Y_i}{Y_{i+1}}$ so that if $X_i \leq X_{i+1}$ then $Y_i \leq Y_{i+1}$ and the asset i is dominated by the asset $i + 1$. Since we have assumed that no assets are dominated we may assume that

$$\frac{X_1}{Y_1} > \frac{X_2}{Y_2} > \dots > \frac{X_k}{Y_k}.$$

If $P = (1 - t, t)$ is a probability vector over the two price relative vectors then according to [36] the portfolio $\vec{b} = (b_1, b_2, \dots, b_n)$ is optimal if and only if

$$(1 - t) \frac{X_i}{b_1 X_1 + \dots + b_k X_k} + t \frac{Y_i}{b_1 Y_1 + \dots + b_k Y_k} \leq 1$$

for all $i \in \{1, 2, \dots, k\}$ with equality if $b_i > 0$. Assume that the portfolio $\vec{b} = \delta_j$ is optimal. Now

$$(1 - t) \frac{X_{j+1}}{X_j} + t \frac{Y_{j+1}}{Y_j} \leq 1$$

is equivalent to

$$t \leq \frac{\frac{X_j}{Y_{j+1}} - \frac{X_{j+1}}{Y_{j+1}}}{\frac{X_j}{Y_j} - \frac{X_{j+1}}{Y_{j+1}}}. \tag{14}$$

Similarly

$$(1 - t) \frac{X_{j-1}}{X_j} + t \frac{Y_{j-1}}{Y_j} \leq 1$$

is equivalent to

$$t \geq \frac{\frac{X_j}{Y_{j-1}} - \frac{X_{j-1}}{Y_{j-1}}}{\frac{X_j}{Y_j} - \frac{X_{j-1}}{Y_{j-1}}}. \tag{15}$$

We have to check that

$$\frac{\frac{X_j}{Y_{j-1}} - \frac{X_{j-1}}{Y_{j-1}}}{\frac{X_j}{Y_j} - \frac{X_{j-1}}{Y_{j-1}}} < \frac{\frac{X_j}{Y_{j+1}} - \frac{X_{j+1}}{Y_{j+1}}}{\frac{X_j}{Y_j} - \frac{X_{j+1}}{Y_{j+1}}},$$

which is equivalent with

$$0 < X_j Y_{j-1} - Y_{j-1} X_{j+1} - Y_j X_{j-1} - (X_j Y_{j+1} - Y_{j+1} X_{j-1} - Y_j X_{j+1}).$$

The right hand side equals the determinant

$$\begin{vmatrix} X_{j+1} - X_{j-1} & X_j - X_{j-1} \\ Y_{j+1} - Y_{j-1} & Y_j - Y_{j-1} \end{vmatrix},$$

which is positive because asset j is not dominated by a portfolio based on asset $j - 1$ and asset $j + 1$.

We see that the portfolio concentrated in asset j is optimal for t in an interval of positive length and the regret between distributions in such an interval will be zero. In particular the regret will not be strict.

Strictness of the regret function is only possible if there are only two assets and if a portfolio concentrated on one of these assets is only optimal for a singular probability measure. According to the formulas for the end points of intervals (14) and (15) this is only possible if the assets are gambling assets. \square

Theorem 1. *If the regret function is strict it equals information divergence, i.e.,*

$$D_G(P, Q) = D(P\|Q). \quad (16)$$

Proof. If the regret function is strict then it is also strict when we restrict to two price relative vectors. Therefore any two price relative vectors are orthogonal gambling assets. If the assets are orthogonal gambling assets we get the type of gambling described by Kelly [35], for gambling equations can easily be derived [36]. \square

5. Sufficiency Conditions

In this section we will introduce some conditions on a regret function. Under some mild conditions they turn out to be equivalent.

Theorem 2. *Let D_F denote a regret function based on a continuous and convex function F defined on the state space of a finite dimensional C^* -algebra. If the state space has at least three orthogonal states then the following conditions are equivalent:*

- *The function F equals entropy times a negative constant plus an affine function.*
- *The regret D_F is proportional to information divergence.*
- *The regret is monotone.*
- *The regret satisfies sufficiency.*
- *The regret is local.*

In the rest of this section we will describe each of these equivalent conditions and prove that they are actually equivalent. The theorems and proofs will be stated so that they hold even for more general state spaces than the ones considered in this paper.

5.1. Entropy and Information Divergence

Definition 4. *Let s denote an element in a state space. The entropy of s is defined as*

$$H(s) = \inf \left(- \sum_{i=1}^n \lambda_i \ln(\lambda_i) \right)$$

where the infimum is taken over all decompositions $s = \sum_{i=1}^n \lambda_i s_i$ of s into pure states s_i .

This definition of the entropy of a state was first given by Uhlmann [38]. Using the fact that entropy is decreasing under majorization we see that the entropy of s is attained at an orthogonal decomposition [13] and we obtain the familiar equation

$$H(s) = -\text{tr}[s \ln(s)].$$

In general this definition of entropy does not provide a concave function on a convex set. For instance, the entropy of points in the square has local maximum in the four different points. A characterization of the convex sets with concave entropy functions is lacking.

Definition 5. If the entropy is a concave function then the regret function D_{-H} is called information divergence.

The information divergence is also called *Kullback–Leibler divergence, relative entropy or quantum relative entropy*. In a C^* -algebra we get

$$\begin{aligned} D_{-H}(s_1, s_2) &= -H(s_1) - (-H(s_2) + \langle s_1 - s_2, -\nabla H(s_2) \rangle) \\ &= H(s_2) - H(s_1) + \langle s_1 - s_2, \nabla H(s_2) \rangle \\ &= \text{tr}[f(s_2)] - \text{tr}[f(s_1)] + \text{tr}[(s_1 - s_2) f'(s_2)] \\ &= \text{tr}[f(s_2) - f(s_1) + (s_1 - s_2) f'(s_2)] \end{aligned}$$

where $f(x) = -x \ln(x)$. Now $f'(x) = -\ln(x) - 1$ so that

$$\begin{aligned} f(s_2) - f(s_1) + (s_1 - s_2) f'(s_2) &= -s_2 \ln(s_2) + s_1 \ln(s_1) + (s_1 - s_2)(-\ln(s_2) - 1) \\ &= s_1(\ln(s_1) - \ln(s_2)) + s_2 - s_1. \end{aligned}$$

Hence

$$D_{-H}(s_1, s_2) = \text{tr}[s_1(\ln(s_1) - \ln(s_2)) + s_2 - s_1].$$

For states s_1, s_2 it reduces to the well-known formula

$$D_{-H}(s_1, s_2) = \text{tr}[s_1 \ln(s_1) - s_1 \ln(s_2)].$$

5.2. Monotonicity

We consider a set \mathcal{T} of maps of the state space into itself. The set \mathcal{T} will be used to represent those transformations that we are able to perform on the state space before we choose a feasible action $a \in \mathcal{A}$. Let $\Phi : \mathcal{S} \curvearrowright \mathcal{S}$ denote a map. Then the dual map Φ^* maps actions into actions and is given by

$$\langle a, \Phi(s) \rangle = \langle \Phi^*(a), s \rangle.$$

Proposition 5 (The principle of lost opportunities). *If Φ^* maps the set of feasible actions \mathcal{A} into itself then*

$$F(\Phi(s)) \leq F(s). \tag{17}$$

Proof. If $a \in \mathcal{A}$ then

$$\begin{aligned} \langle a, \Phi(s) \rangle &= \langle \Phi^*(a), s \rangle \\ &\leq F(s) \end{aligned}$$

because $\Phi^*(a) \in \mathcal{A}$. Inequality (17) follows because $F(\Phi(s)) = \sup_a \langle a, \Phi(s) \rangle$. \square

Corollary 1 (Semi-monotonicity). *Let Φ denote a map of the state space into itself such that Φ^* maps the set of feasible actions \mathcal{A} into itself and let s_2 denote a state that minimizes the function F . If D_F is a Bregman divergence then*

$$D_F(\Phi(s_1), \Phi(s_2)) \leq D_F(s_1, s_2). \tag{18}$$

Proof. Since s_2 minimizes F and F is differentiable we have $\nabla F(s_2) = 0$. Since s_2 minimizes F and $F(\Phi(s_2)) \leq F(s_2)$ we also have that $\Phi(s_2)$ minimizes F and that $\nabla F(\Phi(s_2)) = 0$. Therefore

$$\begin{aligned}
 D_F(\Phi(s_1), \Phi(s_2)) &= F(\Phi(s_1)) - (F(\Phi(s_2)) + \langle \Phi(s_1) - \Phi(s_2), \nabla F(\Phi(s_2)) \rangle) \\
 &= F(\Phi(s_1)) - F(\Phi(s_2)) \\
 &\leq F(s_1) - F(s_2) \\
 &= D_F(s_1, s_2),
 \end{aligned}$$

which proves the inequality. \square

Next we introduce the stronger notion of monotonicity.

Definition 6. Let D_F denote a regret function on the state space S of a finite dimensional C^* -algebra. Then D_F is said to be monotone if

$$D_F(\Phi(s_1), \Phi(s_2)) \leq D_F(s_1, s_2)$$

for any affine map $\Phi : S \rightarrow S$.

Proposition 6. If a regret function D_F based on a convex and continuous function F is monotone then it is a Bregman divergence.

Proof. Assume that D_F is monotone. We have to prove that F is differentiable. Since F is convex it is sufficient to prove that any restriction of F to a line segment is differentiable. Let s_0 and s_1 denote states that are the end points of a line segment. The restriction of F to the line segment is given by the convex and continuous function $f(t) = F((1 - t)s_0 + ts_1)$ so we have to prove that f is differentiable.

If $0 < t_1 < t_2 < 1$ then according to Equation (2) we have

$$D_F((1 - t_2)s_0 + t_2s_1, (1 - t_1)s_0 + t_1s_1) = f(t_2) - (f(t_1) + (t_2 - t_1) \cdot f'_+(t_1))$$

where f'_+ denotes the derivative from the right. A dilation by a factor $r \leq 1$ around s_0 decreases the regret so that

$$r \rightarrow f(r \cdot t_2) - (f(r \cdot t_1) + r \cdot (t_2 - t_1) \cdot f'_+(r \cdot t_1)) \tag{19}$$

is increasing. Since f is convex the function $r \rightarrow f'_+(r \cdot t_1)$ is increasing. Assume that f is not differentiable so that $r \rightarrow f'_+(r \cdot t_1)$ has a positive jump as illustrated on Figure 3.

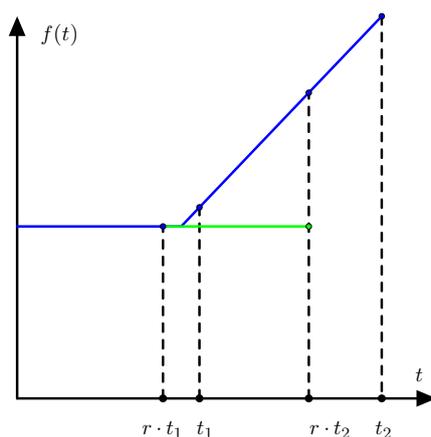


Figure 3. Example of a dilation that increases regret.

This contradicts that the function (19) is increasing. Therefore f'_+ is continuous and f is differentiable. \square

Recently it has been proved that information divergence on a complex Hilbert space is decreasing under positive trace preserving maps [39,40]. Previously this was only known to hold if some extra condition like complete positivity or 2-positivity was assumed [41].

Theorem 3. *Information divergence is monotone under any positive trace preserving map on the states of a finite dimensional C^* -algebra.*

Proof. Any finite dimensional C^* -algebra \mathcal{B} can be embedded in $\mathbb{B}(\mathbb{H})$ and there exist a conditional expectation $\mathbb{E} : \mathbb{B}(\mathbb{H}) \rightarrow \mathcal{B}$. If Φ is a positive trace preserving map of the density matrices of \mathcal{B} into it self then $\Phi \circ \mathbb{E}$ is positive and trace preserving on $\mathbb{B}(\mathbb{H})$. According to Müller-Hermes and Reeb [39] we have

$$D(\Phi \circ \mathbb{E}(s_1) \parallel \Phi \circ \mathbb{E}(s_2)) \leq D(s_1 \parallel s_2)$$

for density matrices in $\mathbb{B}(\mathbb{H})$. In particular this inequality holds for density matrices in \mathcal{B} and for such matrices we have $\mathbb{E}(s_i) = s_i$. \square

5.3. Sufficiency

The notion of sufficiency plays an important role in statistics and related fields. We shall present a definition of sufficiency that is based on [42], but there are a number of other equivalent ways of defining this concept. We refer to [43] where the notion of sufficiency is discussed in great detail.

Definition 7. *Let $(s_\theta)_\theta$ denote a family of states and let Φ denote an affine map $\mathcal{S} \rightarrow \mathcal{T}$ where \mathcal{S} and \mathcal{T} denote state spaces. A recovery map is an affine map $\Psi : \mathcal{T} \rightarrow \mathcal{S}$ such that $\Psi(\Phi(s_\theta)) = s_\theta$. The map Φ is said to be sufficient for $(s_\theta)_\theta$ if Φ has a recovery map.*

Proposition 7. *Assume D_F is a regret function based on a convex and continuous function F and assume that Φ is sufficient for s_1 and s_2 with recovery map Ψ . Assume that both Φ^* and Ψ^* map the set of feasible actions \mathcal{A} into itself. Then*

$$D_F(\Phi(s_1), \Phi(s_2)) = D_F(s_1, s_2).$$

Proof. According to the principle of lest opportunities (Proposition 5) we have

$$\begin{aligned} F(s_2) &= F(\Psi(\Phi(s_2))) \\ &\leq F(\Phi(s_2)) \\ &\leq F(s_2). \end{aligned}$$

Therefore $F(\Phi(s_2)) = F(s_2)$. Let a denote an action that is optimal for s_2 . Then

$$\begin{aligned} F(\Phi(s_2)) &= F(s_2) \\ &= \langle a, s_2 \rangle \\ &= \langle a, \Psi(\Phi(s_2)) \rangle \\ &= \langle \Psi^*(a), \Phi(s_2) \rangle \end{aligned}$$

and we see that $\Psi^*(a)$ is optimal for $\Phi(s_2)$. Now

$$\begin{aligned} D_F(s_1, s_2) &= \inf_a (F(s_1) - \langle a, s_1 \rangle) \\ &= \inf_a (F(s_1) - \langle \Psi^*(a), \Phi(s_1) \rangle) \end{aligned}$$

where the infimum is taken over actions a that are optimal for s_2 . Then

$$\begin{aligned} \inf_a (F(s_1) - \langle \Psi^*(a), \Phi(s_1) \rangle) &\geq \inf_{\tilde{a}} (F(\Phi(s_1)) - \langle \tilde{a}, \Phi(s_1) \rangle) \\ &= D_F(\Phi(s_1), \Phi(s_2)) \end{aligned}$$

so we have $D_F(s_1, s_2) \geq D_F(\Phi(s_1), \Phi(s_2))$. The reverse inequality is proved in the same way. \square

The notion of sufficiency as a property of divergences was introduced in [44]. The crucial idea of restricting the attention to maps of the state space into itself was introduced in [45]. It was shown in [45] that a Bregman divergence on the simplex of distributions on an alphabet that is not binary and satisfies sufficiency equals information divergence up a multiplicative factor. Here we extend the notion of sufficiency from Bregman divergences to regret functions.

Definition 8. Let D_F denote a regret function based on a convex and continuous function F on a state space \mathcal{S} . We say D_F satisfies sufficiency if

$$D_F(\Phi(s_1), \Phi(s_2)) = D_F(s_1, s_2)$$

for any affine map $\mathcal{S} \rightarrow \mathcal{S}$ that is sufficient for (s_1, s_2) .

Proposition 8. Let D_F denote a regret function based on a convex and continuous function F on a state space \mathcal{S} . If the regret function D_F is monotone then it satisfies sufficiency.

Proof. Assume that the regret function D_F is monotone. Let s_1 and s_2 denote two states and let Φ and Ψ denote maps on the state space such that $\Phi(\Psi(s_i)) = s_i$, $i = 1, 2$. Then

$$\begin{aligned} D_F(s_1, s_2) &= D_F(\Phi(\Psi(s_1)), \Phi(\Psi(s_2))) \\ &\leq D_F(\Psi(s_1), \Psi(s_2)) \\ &\leq D_F(s_1, s_2). \end{aligned}$$

Hence $D_F(\Psi(s_1), \Psi(s_2)) = D_F(s_1, s_2)$. \square

Combining the previous results we get that information divergence satisfies sufficiency. Under some conditions there exists an inverse version of Proposition 8 stating that if monotonicity holds with equality then the map is sufficient. In statistics where the state space is a simplex, this result is well established. For density matrices over the complex numbers it has been proved for completely positive maps in [43]. Some new results on this topic can be found in [46].

5.4. Locality

Often it is relevant to use the following weak version of the sufficiency property.

Definition 9. Let D_F denote a regret function based on a convex and continuous function F on a state space \mathcal{S} . The regret function D_F is said to be local if

$$D_F(s_1, t \cdot s_1 + (1-t) \cdot \sigma) = D_F(s_1, t \cdot s_1 + (1-t) \cdot \rho)$$

when the states σ and ρ are orthogonal to s_1 and $t \in]0, 1[$.

Example 5. On a 1-dimensional simplex (an interval) or on the Bloch sphere any regret function D_F is local. The reason is that if σ and ρ are states that are orthogonal to s_1 then $\sigma = \rho$.

Proposition 9. Let D_F denote a regret function based on a convex and continuous function F on a state space S . If the regret function D_F satisfies sufficiency then D_F is local.

Proof. Let σ and ρ be states that are orthogonal to s_1 . Let p denote the projection supporting the state s_0 . Let the maps Φ and Ψ be defined by

$$\begin{aligned} \Phi(s) &= \text{tr}(ps) \cdot s_1 + (1 - \text{tr}(ps)) \cdot \rho, \\ \Psi(s) &= \text{tr}(ps) \cdot s_1 + (1 - \text{tr}(ps)) \cdot \sigma. \end{aligned}$$

Then $\Phi(s_1) = \Psi(s_1) = s_1$ and $\Phi(\sigma) = \rho$ and $\Psi(\rho) = \sigma$. Therefore

$$\begin{aligned} \Phi(t \cdot s_1 + (1 - t) \cdot \sigma) &= t \cdot s_1 + (1 - t) \cdot \rho \\ \Psi(t \cdot s_1 + (1 - t) \cdot \rho) &= t \cdot s_1 + (1 - t) \cdot \sigma \end{aligned}$$

and

$$D_F(s_1, t \cdot s_1 + (1 - t) \cdot \sigma) = D_F(s_1, t \cdot s_1 + (1 - t) \cdot \rho),$$

which proves the Proposition. \square

Theorem 4. Let S be the state space of a C^* -algebra with at least three orthogonal states, and let D_F denote a regret function based on a convex and continuous function F on the state space S . If the regret function D_F is local then it is the Bregman divergence generated by the entropy times a negative constant.

Proof. In the following proof we will assume that the regret function is based on the convex function $F : S \rightarrow \mathbb{R}$. First we will prove that the regret function is a Bregman divergence.

Let K denote the convex hull of a set s_0, s_1, \dots, s_n of orthogonal states. For $x \in [0, 1]$ let g_i denote the function $g_i(x) = D_F(s_i, xs_i + (1 - x)s_{i+1})$. Note that g_i is decreasing and continuous from the left. Let $P = \sum p_i s_i$ and $Q = \sum q_i s_i$ where $p_i, q_i \in]0, 1[$ for all $i = 0, 1, 2, \dots, n$. If F is differentiable in P then locality implies that

$$\begin{aligned} D_F(P, Q) &= \sum p_i D_F(s_i, Q) - \sum p_i D_F(s_i, P) \\ &= \sum p_i g_i(q_i) - \sum p_i g_i(p_i) \\ &= \sum p_i (g_i(q_i) - g_i(p_i)). \end{aligned}$$

Note that $P \rightarrow D_F(P, Q)$ is a convex function and thereby it is continuous. Assume that P_0 is an arbitrary element in K and let $(P_n)_{n \in \mathbb{N}}$ denote a sequence such that $P_n \rightarrow P_0$ for $n \rightarrow \infty$. The sequence $(P_n)_{n \in \mathbb{N}}$ can be chosen so that regret is differentiable in P_n for all $n \in \mathbb{N}$. Further the sequence P_n can be chosen such that $p_{n,i}$ is increasing for all $i \neq j$. Then

$$D_F(P_0, Q) = \sum p_{0,i} (g_i(q_i) - g_i(p_{0,i})) + p_{0,j} g_j(p_{0,j}) - p_{0,j} \lim_{n \rightarrow \infty} g_j(p_{n,j}).$$

Similarly, if the sequence P_n can be chosen such that $p_{n,i}$ is increasing for all $i \neq j, j + 1$ then

$$\begin{aligned} D_F(P_0, Q) &= \sum p_{0,i} (g_i(q_i) - g_i(p_{0,i})) + p_{0,j} g_j(p_{0,j}) - p_{0,j} \lim_{n \rightarrow \infty} g_j(p_{n,j}) \\ &\quad + p_{0,j+1} g_{j+1}(p_{0,j+1}) - p_{0,j+1} \lim_{n \rightarrow \infty} g_{j+1}(p_{n,j+1}), \end{aligned}$$

which implies that $p_{0,j+1} g_{j+1}(p_{0,j+1}) - p_{0,j+1} \lim_{n \rightarrow \infty} g_{j+1}(p_{n,j+1}) = 0$ and that

$$\lim_{n \rightarrow \infty} g_{j+1}(p_{n,j+1}) = g_{j+1}(p_{0,j+1})$$

for all j . Therefore

$$D_F(P_0, Q) = \sum p_{0,i} (g_i(q_i) - g_i(p_{0,i})) \tag{20}$$

for all P_0, Q in the interior of K . In the following calculations we will assume that the distributions lie in the interior of K . The validity of the Bregman identity (5) follows directly from Equation (20) implying that D_F is a Bregman divergence.

As a function of Q the regret is minimal when $Q = P$. In the following calculations we write $x = p_i, z = p_j, y = q_i$, and $w = q_j$. If $p_\ell = q_\ell$ for $\ell \neq i, j$ then non-negativity of regret can be written as

$$x (g_i(y) - g_i(x)) + z (g_j(w) - g_j(z)) \geq 0$$

and we note that this inequality should hold as long as $x + z = y + w \leq 1$. Permutation of i and j leads to the inequality

$$x (g_j(y) - g_j(x)) + z (g_i(w) - g_i(z)) \geq 0$$

that implies

$$x (g_{ij}(y) - g_{ij}(x)) + z (g_{ij}(w) - g_{ij}(z)) \geq 0 \tag{21}$$

where $g_{ij} = \frac{g_i + g_j}{2}$.

Assume that $x = z = \frac{y+w}{2}$ in Inequality (21). Then

$$\begin{aligned} x (g_{ij}(y) - g_{ij}(x)) + x (g_{ij}(w) - g_{ij}(x)) &\geq 0 \\ g_{ij}(y) - g_{ij}(x) + g_{ij}(w) - g_{ij}(x) &\geq 0 \\ \frac{g_{ij}(y) + g_{ij}(w)}{2} &\geq g_{ij}(x) \end{aligned}$$

so that g_{ij} is mid-point convex, which for a measurable function implies convexity. Therefore g_{ij} is differentiable from left and right.

If $y = w$ and $x = y + \epsilon$ and $z = y - \epsilon$ then we have

$$(y + \epsilon) (g_{ij}(y) - g_{ij}(y + \epsilon)) + (y - \epsilon) (g_{ij}(y) - g_{ij}(y - \epsilon)) \geq 0$$

with equality when $\epsilon = 0$. We differentiate with respect to ϵ from right.

$$(g_{ij}(y) - g_{ij}(y + \epsilon)) + (y + \epsilon) (-g'_{ij+}(y + \epsilon)) - (g_{ij}(y) - g_{ij}(y - \epsilon)) + (y - \epsilon) (g'_{ij-}(y - \epsilon)),$$

which is positive for $\epsilon = 0$ so that

$$-y \cdot g'_{ij+}(y) + y \cdot g'_{ij-}(y) \geq 0 \tag{22}$$

$$y \cdot g'_{ij-}(y) \geq y \cdot g'_{ij+}(y). \tag{23}$$

Since g_{ij} is convex we have $g'_{ij-}(y) \leq g'_{ij+}(y)$ which in combination with Inequality (23) implies that $g'_{ij-}(y) = g'_{ij+}(y)$ so that g_{ij} is differentiable. Since $g_i = g_{ij} + g_{ik} - g_{jk}$ the function g_i is also differentiable.

As a function of Q the Bregman divergence $D_F(P, Q)$ has a minimum at $Q = P$ under the condition $\sum q_i = 1$. Since the functions g_i are differentiable we can characterize this minimum using Lagrange multipliers. We have

$$\frac{\partial}{\partial q_i} D_F(P, Q) = p_i g'_i(q_i)$$

and

$$\frac{\partial}{\partial q_i} D_F(P, Q)|_{Q=P} = p_i \cdot g'_i(p_i).$$

Further $\frac{\partial}{\partial q_i} \sum q_i = 1$ so there exist a constant c_K such that $p_i \cdot g'_i(p_i) = c_K$. Hence $g'_i(p_i) = \frac{c_K}{p_i}$ so that $g_i(p_i) = c_K \cdot \ln(p_i) + m_i$ for some constant m_i .

Now we get

$$\begin{aligned} D_F(P, Q) &= \sum p_i (g_i(q_i) - g_i(p_i)) \\ &= \sum p_i ((c_K \cdot \ln(q_i) + m_i) - (c_K \cdot \ln(p_i) + m_i)) \\ &= -c_K \cdot \sum p_i \ln \frac{p_i}{q_i} \\ &= D_{-c_K \cdot H}(P, Q). \end{aligned}$$

Therefore, an affine function exists, defined by K such that

$$F|_K(P) = -c_K \cdot H|_K(P) + g_K \tag{24}$$

for all P in the interior of K . Since H_K is continuous on K Equation (24) holds for any $P \in K$. If each of the sets K and L is a simplex and $x \in K \cap L$ then

$$-c_K \cdot H|_K(x) + g_K(x) = -c_L \cdot H|_L(x) + g_L(x)$$

so that

$$(c_L - c_K) \cdot H|_K(x) = g_L(x) - g_K(x).$$

If $K \cap L$ has dimension greater than zero then the right hand side is affine so the left hand side is affine, which is only possible when $c_K = c_L$. Therefore we also have $g_L(x) = g_K(x)$ for all $x \in K \cap L$. Therefore the functions g_K can be extended to a single affine function on the whole of \mathcal{S} . \square

6. Applications

6.1. Information Theory

If only integer values of a code-length function ℓ are allowed then there are only finitely many actions that are not dominated. Therefore the function F given by

$$F(P) = -\min_{\ell} \sum \ell(a) \cdot p_a$$

is piece-wise linear. In particular F is not differentiable so that the regret is not a Bregman divergence and cannot be monotone according to Proposition 6. In information theory monotonicity of a divergence function is closely related to the *data processing inequality* and since the data processing inequality is one of the most important tools for deriving inequalities in information theory we need to modify our notion of code-length function in order to achieve a data processing inequality.

We now formulate a version of Kraft's inequality that allows the code length function to be non-integer valued.

Theorem 5. *Let $\ell : \mathbb{A} \rightarrow \mathbb{R}$ be a function. Then the function ℓ satisfies Kraft's inequality (6) if and only if for all $\varepsilon > 0$ there exists an integer n and a uniquely decodable fixed-to-variable length block code $\kappa : \mathbb{A}^n \rightarrow \mathbb{B}^*$ such that*

$$\left| \bar{\ell}_\kappa(a^n) - \frac{1}{n} \sum_{i=1}^n \ell(a_i) \right| \leq \varepsilon$$

where $\bar{\ell}_\kappa(a^n)$ denotes the length $\ell_\kappa(a^n)$ divided by n . The uniquely decodable block code can be chosen to be prefix free.

Proof. Assume that ℓ satisfies Kraft’s inequality. Then

$$\sum_{a_1 a_2 \dots a_n \in \mathbb{A}^n} \beta^{-\sum_{i=1}^n \ell(a_i)} = \left(\sum_{a \in \mathbb{A}} \beta^{-\ell(a)} \right)^n \leq 1^n = 1.$$

Therefore the function $\tilde{\ell} : \mathbb{A}^n \rightarrow \mathbb{N}$ given by

$$\tilde{\ell}(a_1 a_2 \dots a_n) = \left\lceil \sum_{i=1}^n \ell(a_i) \right\rceil$$

is integer valued and satisfies Kraft’s inequality (6) and there exists a prefix-free code $\kappa : \mathbb{A}^n \rightarrow \{0, 1\}^*$ such that $\ell_\kappa(a_1 a_2 \dots a_n) = \tilde{\ell}(a_1 a_2 \dots a_n)$. Therefore

$$\left| \bar{\ell}_\kappa(a_1 a_2 \dots a_n) - \frac{1}{n} \sum_{i=1}^n \ell(a_i) \right| = \frac{1}{n} \left| \left\lceil \sum_{i=1}^n \ell(a_i) \right\rceil - \sum_{i=1}^n \ell(a_i) \right| \leq \frac{1}{n}$$

so for any $\varepsilon > 0$ choose n such that $1/n \leq \varepsilon$.

Assume that for all $\varepsilon > 0$ there exists a uniquely decodable fixed-to-variable length code $\kappa : \mathbb{A}^n \rightarrow \{0, 1\}^*$ such that

$$\left| \bar{\ell}_\kappa(a_1 a_2 \dots a_n) - \frac{1}{n} \sum_{i=1}^n \ell(a_i) \right| \leq \varepsilon$$

for all strings $a_1 a_2 \dots a_n \in \mathbb{A}^n$. Then $n \bar{\ell}_\kappa(a_1 a_2 \dots a_n)$ satisfies Kraft’s Inequality (6) and

$$\begin{aligned} \left(\sum_{a \in \mathbb{A}} \beta^{-\ell(a)} \right)^n &= \sum_{a_1 a_2 \dots a_n \in \mathbb{A}^n} \beta^{-\sum_{i=1}^n \ell(a_i)} \\ &\leq \sum_{a_1 a_2 \dots a_n \in \mathbb{A}^n} \beta^{-n(\bar{\ell}_\kappa(a_1 a_2 \dots a_n) - \varepsilon)} \\ &= \beta^{n\varepsilon} \sum_{a_1 a_2 \dots a_n \in \mathbb{A}^n} \beta^{-n \bar{\ell}_\kappa(a_1 a_2 \dots a_n)} \\ &\leq \beta^{n\varepsilon}. \end{aligned}$$

Therefore $\sum_{a \in \mathbb{A}} \beta^{-\ell(a)} \leq \beta^\varepsilon$ for all $\varepsilon > 0$ and the result is obtained. \square

Like in Bayesian statistics we focus on finite sequences. Contrary to Bayesian statistics we should always consider a finite sequence as a prefix of *longer finite* sequences. Contrary to frequential statistics we do not have to consider a finite sequence as a prefix of an *infinite* sequence.

If we minimize the mean code-length over functions that satisfy Kraft’s inequality (6), but without an integer constraint the code-length should be $\ell(a) = -\log_\beta(p_a)$ and the function F is given by

$$F(P) = \sum_a p_a \cdot \log_\beta(p_a).$$

The function F is proportional to the Shannon entropy and the (negative) proportionality factor is determined by the size of the output alphabet.

In lossy source coding and rate distortion theory it is important to choose a distortion function with tractable properties. The notion of sufficiency for divergence functions was introduced in [44] in order to characterize such tractable distortions functions. In this paper the main result was that sufficiency together with properties related to Bregman divergence lead directly to the information bottleneck method introduced by N. Tishby [47]. Logarithmic loss has also been studied for lossy compression in [48].

6.2. Statistics

In statistics one is often interested in scoring rules that are local, which means a scoring rule where the payoff only depends on the probability of the observed value and not on the predicted distribution over unobserved values. The notion of locality has recently been extended by Dawid, Lauritzen and Parry [49], but here we shall focus on the original definition. The basic result is that the only local strictly proper scoring rule is logarithmic score that was proved by Bernardo under the assumption that scoring rule is given by a smooth function [50].

Definition 10. A local strictly proper scoring rule is a scoring rule of the form $f(x, Q) = g(Q(x))$.

Theorem 6. On a finite space a local strictly proper scoring rule is given by a local regret function.

Proof. The regret function of a local strictly proper scoring rule is given by

$$D(P, Q) = \sum_x P(x) (g(P(x)) - g(Q(x))).$$

If $Q = (1 - t)P + tQ_i$ and P and Q are mutually singular then

$$\begin{aligned} D(P, Q) &= \sum_x P(x) (g(P(x)) - g((1 - t)P(x) + tQ_i(x))) \\ &= \sum_x P(x) (g(P(x)) - g((1 - t)P(x) + 0)) \end{aligned}$$

and we see that the regret does not depend on Q_i because Q_i vanish on the support of P . Therefore the regret function is local. \square

Corollary 2. On a finite space with at least three elements a local strictly proper scoring rule is given by a function g of the form $g(x) = a \cdot \ln(x) + b$ for some constants a and b .

Also the notion of sufficiency plays an important role in statistics. Here we will restrict the discussion to 1-dimensional exponential families. A natural exponential family is a family of probability distributions of the form

$$\frac{dP_\beta}{dQ} = \frac{\exp(\beta x)}{Z(\beta)}$$

where Q is a reference measure on the real numbers and Z is the moment generating function given by $Z(\beta) = \int \exp(\beta x) dQx$. Then $x_1^n \rightarrow x_1 + x_2 + \dots + x_n$ is a sufficient statistic for the family $(P_\beta^{\otimes n})_\beta$.

Example 6. In a Bernoulli model a sequence $x_1^n \in \{0, 1\}^n$ is predicted with probability

$$\prod_{i=1}^n p^{x_i} (1 - p)^{1 - x_i} = \exp\left(\left(\sum_{i=1}^n x_i\right) \ln(p) + \left(n - \sum_{i=1}^n x_i\right) \ln(1 - p)\right).$$

The function $x_1^n \rightarrow x_1 + x_2 + \dots + x_n$ induces a sufficient map Φ from probability distributions on $\{0, 1\}^n$ to probability distributions on $\{0, 1, 2, \dots, n\}$. The reverse map maps a measure concentrated in $j \in \{0, 1, 2, \dots, n\}$ into a uniform distributions over sequences $x_1^n \in \{0, 1\}^n$ that satisfy $\sum_{i=1}^n x_i = j$.

The mean value of P_β is

$$\int x \cdot \frac{\exp(\beta x)}{Z(\beta)} dQx.$$

The set of possible mean values is called the mean value range and is an interval. Let P^μ denote the element in the exponential family with mean value μ . Then a Bregman divergence on the mean

value range is defined by $D(\lambda, \mu) = D(P^\lambda \| P^\mu)$. Note that the mapping $\mu \rightarrow P^\mu$ is not affine so the Bregman divergence $D(\lambda, \mu)$ will in general not be given by the formula for information divergence with the family of binomial distributions as the only exception. Nevertheless the Bregman divergence $D(\lambda, \mu)$ encodes important information about the exponential family. In statistics it is common to use squared Euclidean distance as distortion measure, but often it is better to use the Bregman divergence $D(\lambda, \mu)$ as a distortion measure. Note that $D(\lambda, \mu)$ is only proportional to squared Euclidean distance for the Gaussian location family.

Example 7. An exponential distribution has density

$$f_\lambda(x) = \begin{cases} \frac{1}{\lambda} \exp(-\frac{x}{\lambda}), & \text{for } x \geq 0; \\ 0, & \text{else.} \end{cases}$$

This leads to a Bregman divergence on the interval $[0; \infty[$ given by

$$\begin{aligned} \int_0^\infty f_\lambda(x) \ln\left(\frac{f_\lambda(x)}{f_\mu(x)}\right) dx &= \frac{\lambda}{\mu} - 1 - \ln\left(\frac{\lambda}{\mu}\right) \\ &= D_{-\ln}(\lambda, \mu). \end{aligned}$$

This Bregman divergence is called the Isakura-Saito distance. The Isakura-Saito distance is defined as an unbounded set so our previous results cannot be applied. Affine bijections on $[0; \infty[$ have the form $\Phi(x) = k \cdot x$ for some constant $k > 0$. The Isakura-Saito distance obviously satisfy sufficiency for such maps and it is a simple exercise to check that the Isakura-Saito distance is the only Bregman divergence on $[0, \infty[$ that satisfies sufficiency. Any affine map $[0; \infty[\rightarrow [0; \infty[$ is composed of a map $x \rightarrow k \cdot x$ where $k \geq 0$ and a right translation $x \rightarrow x + t$ where $t \geq 0$. The Itakura-Saito distance decreases under right translations because

$$\begin{aligned} \frac{\partial}{\partial t} D_{-\ln}(\lambda + t, \mu + t) &= \frac{\partial}{\partial t} \left(\frac{\lambda + t}{\mu + t} - 1 - \ln\left(\frac{\lambda + t}{\mu + t}\right) \right) \\ &= \frac{(\mu + t) - (\lambda + t)}{(\mu + t)^2} - \frac{1}{\lambda + t} + \frac{1}{\mu + t} \\ &= -\frac{(\lambda - \mu)^2}{(\lambda + t)(\mu + t)^2} \leq 0. \end{aligned}$$

Thus the Isakura-Saito distance is monotone.

Both sufficiency and the Bregman identity are closely related to inference rules. In [51] I. Csiszár explained why information divergence is the only divergence function on the cone of positive measures that lead to certain tractable inference rules. One should observe that his inference rules are closely related to sufficiency and the Bregman identity, and the present paper may be viewed as a generalization of these results of I. Csiszár.

In the minimum description length approach to statistics [9] it is common to minimize the maximal regret of the model where the maximum is taken over all possible data sequences. For most exponential families this approach is computationally difficult and may cause problems with normalization of the optimal distribution over the parameter. In general this approach will also depend on the length of the data sequence in a way that is not transitive. That means that one cannot analyze a subsequence before the length of the whole data sequence is known. In [52] it was proved that for one dimensional exponential family there are essentially three exponential families where these problems are avoided. The exponential families are the Gaussian location family, the Gamma distributions, and the Tweedie distributions of order $3/2$. The statistical analysis of the Gaussian location family reduces to minimizing the sum of squares. Similarly, the Gamma distributions can be analyzed using the Isakura-Saito distance (or information divergence), but the Tweedie family of order $3/2$ is an exotic object that has

not been analyzed in similar detail. For exponential models in two or more dimensions similar results are not known, but in general one should expect that most models are complicated to analyze exactly while certain models simplify due to some type of inner symmetry of the model.

6.3. Statistical Mechanics

Statistical mechanics can be stated based on classical mechanics or quantum mechanics. For our purpose this makes no difference because our theorems are valid for both classical systems and quantum systems.

As we have seen before

$$Ex = kT_0 \cdot D(s \| s_0). \quad (25)$$

Our general results for Bregman divergences imply that the Bregman divergence based on this exergy satisfies

$$D_{Ex}(s_1, s_2) = kT_0 \cdot D(s_1 \| s_2).$$

Therefore

$$D_{Ex}(\Phi(s_1), \Phi(s_2)) = D_{Ex}(s_1, s_2)$$

for any map that is sufficient for $\{s_1, s_2\}$. The equality holds for any regret function that is reversible and conserves the state that is in equilibrium with the environment. Since a different temperature of the environment leads to a different state that is in equilibrium the equality holds for any reversible map that leave some equilibrium state invariant. We see that $D_{Ex}(s_1, s_2)$ is uniquely determined as long as there exists a sufficiently large set of maps that are reversible.

In this exposition we have made some short-cuts. First of all we did not derive Equation (25). In particular the notion of temperature was used without discussion. Secondly we identified the internal energy with the mean value of the Hamiltonian and identified the thermodynamic entropy with k times the Shannon entropy. Finally, in the argument above we need to verify in all details that the set of reversible maps is sufficiently large to determine the regret function. For classical thermodynamics a comprehensive exposition was made by Lieb and Yngvason [53,54]. In their exposition randomness was not taken into account. With the present framework it is also possible to handle randomness so that one can make a bridge between thermodynamics and statistical mechanics. The approach of Lieb and Yngvason was recently improved by C. Marletto [55] uses the formalism of constructor theory to derive results. The basic idea in constructor theory is to distinguish between possible and impossible transformation in a way that is closely related to the ideas presented in this paper. A detailed exposition of such relations will be given in a future paper.

According to Equation (25) any bit of information can be converted into an amount of energy! One may ask how this is related to the mixing paradox (a special case of Gibbs' paradox). Consider a container divided by a wall with a blue and a yellow gas on each side of the wall as illustrated in Figure 4. The question is how much energy can be extracted by mixing the blue and the yellow gas?

We loose one bit of information about each molecule by mixing the blue and the green gas, but if the color is the *only difference* no energy can be extracted. This seems to be in conflict with Equation (25), but in this case different states cannot be converted into each other by reversible processes. For instance one cannot convert the blue gas into the yellow gas. To get around this problem one can restrict the set of preparations and one can restrict the set of measurements. For instance one may simply ignore measurements of the color of the gas. What should be taken into account and what should be ignored, can only be answered by an experienced physicist. Formally this solves the mixing paradox, but from a practical point of view nothing has been solved. If for instance the molecules in one of the gases are much larger than the molecules in the other gas then a semi-permeable membrane can be used to create an osmotic pressure that can be used to extract some energy. It is still an open question as to which differences in properties of the two gases that can be used to extract energy.

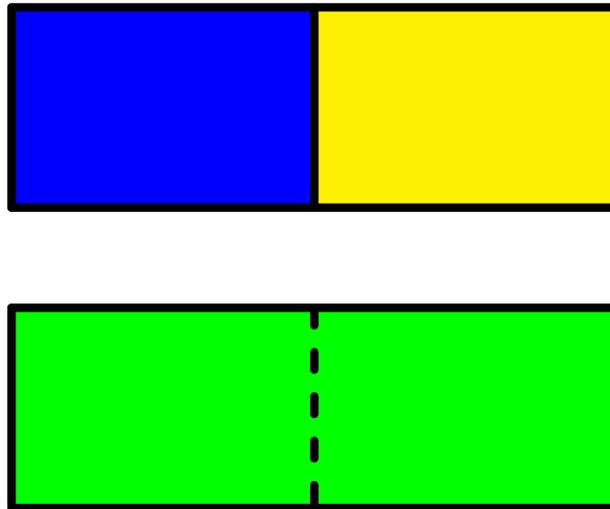


Figure 4. Mixing of a blue and a yellow gas.

6.4. Monotone Regret for Portfolios

We know that in general a local regret function on a state space with at least three orthogonal states is proportional to information divergence. In portfolio theory we get the stronger result that monotonicity implies that we are in the situation of gambling introduced by Kelly [35].

Theorem 7. *Assume that none of the assets are dominated by a portfolio of other assets. If the regret function $D_G(P, Q)$ given by (13) is monotone then the regret function equals information divergence and the measures P and Q are supported by k distinct price relative vectors of the form $(o_1, 0, 0, \dots, 0)$, $(0, o_2, 0, \dots, 0)$, until $(0, 0, \dots, o_k)$.*

Proof. If there are more than three price relative vectors then a monotone regret function is always proportional to information divergence which is a strict regret function. Therefore we may assume that there are only two price relative vectors. Assume that the regret function is not strict. Then the function G defined by (12) is not strictly convex. Assume that $D_G(P, Q) = 0$ so that G is affine between P and Q . Let Φ be a contraction around one of the end points of intersection between the state space and the line through P and Q . Then monotonicity implies that $D_G(\Phi(P), \Phi(Q)) = 0$ so that G is affine on the line between $\Phi(P)$ and $\Phi(Q)$. This holds for contractions around any point. Therefore G is affine on the whole state space which implies that there is a single portfolio that dominates all assets. Such a portfolio must be supported on a single asset. Therefore monotonicity implies that if two assets are not dominated then the regret function is strict and according to Theorem 1 we have already proved that a strict regret function in portfolio theory is proportional to information divergence. \square

Example 8. *If the regret function divergence is monotone and one of the assets is the safe asset then there exists a portfolio \vec{b} such that $b_i \cdot o_i \geq 1$ for all i . Equivalently $b_i \geq o_i^{-1}$ which is possible if and only if $\sum o_i^{-1} \leq 1$. One says that the gamble is fair if $\sum o_i^{-1} = 1$. If the gamble is super-fair, i.e., $\sum o_i^{-1} < 1$, then the portfolio $b_i = o_i^{-1} / \sum o_i^{-1}$ gives a price relative equal to $(\sum o_i^{-1})^{-1} > 1$ independently of what happens, which is a Dutch book.*

Corollary 3. *In portfolio theory the regret function $D_G(P, Q)$ given by (13) is monotone if and only if it is strict.*

Proof. We use that in portfolio theory the regret function is monotone if and only if it is proportional to information. \square

7. Concluding Remarks

Convexity of a Bregman divergence is an important property that was first studied systematically in [56] and extended from probability distributions to matrices in [57]. In [58] it was proved that if f is a function such that the Bregman divergence based on $\text{tr}(f(\rho))$ is monotone on any (simple) C^* -algebra then the Bregman divergence is jointly convex. As we have seen monotonicity implies that the Bregman divergence must be proportional to information divergence, which is jointly convex in both arguments. We also see that in general joint convexity is not a sufficient condition for monotonicity, but in the case where the state space has only two orthogonal states it is not known if joint convexity of a Bregman divergence is sufficient to conclude that the Bregman divergence is monotone.

One should note that the type of optimization presented in this paper is closely related to a game theoretic model developed by F. Topsøe [59,60]. In his game theoretic model he needed what he called the *perfect match principle*. Using the terminology presented in this paper the perfect match principle states that the regret function is a strict Bregman divergence. As we have seen the perfect match principle is only fulfilled in portfolio theory if all the assets are gambling assets. Therefore, the theory of F. Topsøe can only be used to describe gambling while our optimization model can describe general portfolio theory and our sufficient conditions can explain exactly when our general model equals gambling. The formalism that has been developed in this paper is also closely related to constructor theory [61], but a discussion will be postponed to another article.

The original paper of Kullback and Leibler [1] was called “On Information and Sufficiency”. In the present paper, we have made the relation between information divergence and the notion of sufficiency more explicit. The results presented in this paper are closely related to the result that a divergence that is both an f -divergence and a Bregman divergence is proportional to information divergence (see [44] or [62] and references therein). All f -divergences satisfy a sufficiency condition, which is the reason why this class of divergences has played such a prominent role in the study of the relation between information theory and statistics. One major question has been to find reasons for choosing between the different f -divergences. For instance f -divergences of power type (often called Tsallis divergences or Cressie-Read divergences) are popular [5], but there are surprisingly few papers that can point at a single value of the power α that is optimal for a certain problem except if this value is 1. In this paper we have started with Bregman divergences because each optimization problem comes with a specific Bregman divergence. Often it is possible to specify a Bregman divergence for an optimization problem and only in some of the cases this Bregman divergence is proportional to information divergence.

The idea of sufficiency has different relevance in different applications, but in all cases information divergence prove to be the quantity that convert the general notion of sufficiency into a number. In information theory information divergence appear as a consequence of Kraft’s inequality. For code length functions of integer length we get functions that are piecewise linear. Only if we are interested in extend-able sequences we get a regret function that satisfies a data processing inequality. In this sense information theory is a theory of extend-able sequences. For scoring functions in statistics the notion of locality is important. These applications do not refer to sequences. Similarly the notion of sufficiency that plays a major role in statistics, does not refer to sequences. Both sufficiency and locality imply that regret is proportional to information divergence, but these reasons are different from the reasons why information divergence is used in information theory. Our description of statistical mechanics does not go into technical details, but the main point is that the many symmetries in terms of reversible maps form a set of maps so large that our result on invariance of regret under sufficient maps applies. In this sense statistical mechanics and statistics both apply information divergence for reasons related to sufficiency. For portfolio theory the story is different. In most cases one has to apply the general

theory of Bregman divergences because we deal with an optimization problem. The general Bregman divergences only reduce to information divergence when the assets are gambling assets.

Often one talks about applications of information theory in statistics, statistical mechanics and portfolio theory. In this paper we have argued that information theory is mainly a theory of sequences, while some problems in statistics and statistical mechanics are also relevant without reference to sequences. It would be more correct to say that convex optimization has various application such as information theory, statistics, statistical mechanics, and portfolio theory and that certain conditions related to sufficiency lead to the same type of quantities in all these applications.

Acknowledgments: The author want to thank Prasad Santhanam for inviting me to the Electrical Engineering Department, University of Hawai'i at Mānoa, where many of the ideas presented in this paper were developed. I also want to thank Alexander Müller-Hermes, Frank Hansen, and Flemming Topsøe for stimulating discussions and correspondence. Finally, I want to thank the reviewers for their valuable comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Kullback, S.; Leibler, R. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
2. Jaynes, E.T. Information Theory and Statistical Mechanics, I. *Phys. Rev.* **1957**, *106*, 620–630.
3. Jaynes, E.T. Information Theory and Statistical Mechanics, II. *Phys. Rev.* **1957**, *108*, 171–190.
4. Jaynes, E.T. Clearing up mysteries—The original goal. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer: Dordrecht, The Netherlands, 1989.
5. Liese, F.; Vajda, I. *Convex Statistical Distances*; Teubner: Leipzig, Germany, 1987.
6. Barron, A.R.; Rissanen, J.; Yu, B. The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760.
7. Csiszár, I.; Shields, P. *Information Theory and Statistics: A Tutorial*; Foundations and Trends in Communications and Information Theory; Now Publishers Inc.: Delft, The Netherlands, 2004.
8. Grünwald, P.D.; Dawid, A.P. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Ann. Math. Stat.* **2004**, *32*, 1367–1433.
9. Grünwald, P. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
10. Holevo, A.S. *Probabilistic and Statistical Aspects of Quantum Theory*; North-Holland Series in Statistics and Probability; North-Holland: Amsterdam, The Netherlands, 1982; Volume 1.
11. Krumm, M.; Barnum, H.; Barrett, J.; Müller, M. Thermodynamics and the structure of quantum theory. *arXiv* **2016**, arXiv:1608.04461.
12. Barnum, H.; Müller, M.P.; Ududec, C. Higher-order interference and single-system postulates characterizing quantum theory. *New J. Phys.* **2014**, *16*, 123029.
13. Harremoës, P. Maximum Entropy and Sufficiency. *arXiv* **2016**, arXiv:1607.02259.
14. Harremoës, P. Quantum information on Spectral Sets. *arXiv* **2017**, arXiv:1701.06688.
15. Barnum, H.; Lee, C.M.; Scandolo, C.M.; Selby, J.H. Ruling out higher-order interference from purity principles. *arXiv* **2017**, arXiv:1704.05106.
16. Servage, L.J. The Theory of Statistical Decision. *J. Am. Stat. Assoc.* **1951**, *46*, 55–67.
17. Bell, D.E. Regret in decision making under uncertainty. *Oper. Res.* **1982**, *30*, 961–981.
18. Fishburn, P.C. *The Foundations of Expected Utility*; Springer: Berlin/Heidelberg, Germany, 1982.
19. Loomes, G.; Sugden, R. Regret theory: An alternative theory of rational choice under uncertainty. *Econ. J.* **1982**, *92*, 805–824.
20. Bikhchandani, S.; Segal, U. Transitive regret. *Theor. Econ.* **2011**, *6*, 95–108.
21. Kiwiel, K.C. Proximal Minimization Methods with Generalized Bregman Functions. *SIAM J. Control Optim.* **1997**, *35*, 1142–1168.
22. Kiwiel, K.C. Free-steering Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints. *Math. Oper. Res.* **1997**, *22*, 326–349.
23. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1970.
24. Hendrickson, A.D.; Buehler, R.J. Proper scores for probability forecasters. *Ann. Math. Stat.* **1971**, *42*, 1916–1921.

25. Rao, C.R.; Nayak, T.K. Cross Entropy, Dissimilarity Measures, and Characterizations of Quadratic Entropy. *IEEE Trans. Inf. Theory* **1985**, *31*, 589–593.
26. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
27. Kraft, L.G. A Device for Quantizing, Grouping and Coding Amplitude Modulated Pulses. Master's Thesis, Department of Electrical Engineering, MIT University, Cambridge, MA, USA, 1949.
28. Han, T.S.; Kobayashi, K. *Mathematics of Information and Coding*; Translations of Mathematical Monographs; American Mathematical Society: Providence, RI, USA, 2002; Volume 203.
29. De Finetti, B. *Theory of Probability*; Wiley: Hoboken, NJ, USA, 1974.
30. McCarthy, J. Measures of the value of information. *Proc. Natl. Acad. Sci. USA* **1956**, *42*, 654–655.
31. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
32. Ovcharov, E.Y. Proper Scoring Rules and Bregman Divergences. *arXiv* **2015**, arXiv:1502.01178.
33. Gundersen, T. *An Introduction to the Concept of Exergy and Energy Quality*; Lecture notes; Norwegian University of Science and Technology: Trondheim, Norway, 2011.
34. Harremoës, P. *Time and Conditional Independence*; IMFUFA-Tekst; IMFUFA Roskilde University: Roskilde, Denmark, 1993; Volume 255.
35. Kelly, J.L. A New Interpretation of Information Rate. *Bell Syst. Tech. J.* **1956**, *35*, 917–926.
36. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 1991.
37. Cover, T.M. Universal portfolios. *Math. Finance* **1991**, *1*, 1–29.
38. Uhlmann, A. On the Shannon Entropy and Related Functionals on Convex Sets. *Rep. Math. Phys.* **1970**, *1*, 147–159.
39. Müller-Hermes, A.; Reeb, D. Monotonicity of the Quantum Relative Entropy under Positive Maps. *Annales Henri Poincaré* **2017**, *18*, 1777–1788.
40. Christandl, M.; Müller-Hermes, A. Relative Entropy Bounds on Quantum, Private and Repeater Capacities. *arXiv* **2016**, arXiv:1604.03448.
41. Petz, D. Monotonicity of Quantum Relative Entropy Revisited. *Rev. Math. Phys.* **2003**, *15*, 79–91.
42. Petz, D. Sufficiency of Channels over von Neumann algebras. *Q. J. Math. Oxf.* **1988**, *39*, 97–108.
43. Jenčová, A.; Petz, D. Sufficiency in quantum statistical inference. *Commun. Math. Phys.* **2006**, *263*, 259–276.
44. Harremoës, P.; Tishby, N. The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. In Proceedings of the IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–571.
45. Jiao, J.; Courtade, T.A.; No, A.; Venkat, K.; Weissman, T. Information Measures: The Curious Case of the Binary Alphabet. *IEEE Trans. Inf. Theory* **2014**, *60*, 7616–7626.
46. Jenčová, A. Preservation of a quantum Rényi relative entropy implies existence of a recovery map. *J. Phys. A Math. Theor.* **2017**, *50*, 085303.
47. Tishby, N.; Pereira, F.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Urbana, Illinois, USA, 22–24 September 1999; pp. 368–377.
48. No, A.; Weissman, T. Universality of logarithmic loss in lossy compression. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hongkong, China, 14–19 June 2015; pp. 2166–2170.
49. Dawid, A.P.; Lauritzen, S.; Perry, M. Proper local scoring rules on discrete sample spaces. *Ann. Stat.* **2012**, *40*, 593–603.
50. Bernardo, J.M. Expected Information as Expected Utility. *Ann. Stat.* **1978**, *7*, 686–690.
51. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032–2066.
52. Bartlett, P.; Grünwald, P.; Harremoës, P.; Hedayati, F.; Kotlowski, W. Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families. In Proceedings of the Conference on Learning Theory (COLT 2013), Princeton, NJ, USA, 12–14 June 2013; p. 23.
53. Lieb, E.; Yngvason, J. A Guide to Entropy and the Second Law of Thermodynamics. *Not. AMS* **1998**, *45*, 571–581.

54. Lieb, E.; Yngvason, J. The Mathematics of the Second Law of Thermodynamics. In *Visions in Mathematics*; Alon, N., Bourgain, J., Connes, A., Gromov, M., Milman, V., Eds.; Birkhäuser: Basel, Switzerland, 2010; pp. 334–358.
55. Marletto, C. Constructor Theory of Thermodynamics. *arXiv* **2016**, arXiv:1608.02625.
56. Bauschke, H.H.; Borwein, J.M. Joint and Separate Convexity of the Bregman Distance. In *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*; Dan Butnariu, Y.C.; Reich, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2001; Volume 8, pp. 23 – 36.
57. Hansen, F.; Zhang, Z. Characterisation of Matrix Entropies. *Lett. Math. Phys.* **2015**, *105*, 1399–1411.
58. Pitrik, J.; Virosztek, D. On the Joint Convexity of the Bregman Divergence of Matrices. *Lett. Math. Phys.* **2015**, *105*, 675–692.
59. Topsøe, F. Game theoretical optimization inspired by information theory. *J. Glob. Optim.* **2008**, *43*, 553–564.
60. Topsøe, F. Cognition and Inference in an Abstract Setting. In Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2011), Helsinki, Finland, 7–10 August 2011; pp. 67–70.
61. Deutch, D.; Marletto, C. Constructor theory of information. *Proc. R. Soc. A* **2014**, *471*, 20140540.
62. Amari, S.I. α -Divergence Is Unique, Belonging to Both f -Divergence and Bregman Divergence Classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.



© 2017 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).