Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics

Flemming Topsøe

University of Copenhagen, Institute of Mathematical Sciences
Universitetsparken 5, 2100 Copenhagen, Denmark

Abstract. For a wide range of entropy measures, easy calculation of equilibria is possible using a principle of *Game Theoretical Equilibrium* related to Jaynes *Maximum Entropy Principle*. This follows previous work of the author and relates to Naudts [1], [2], and, partly, Abe and Bagci [3].

Keywords: Complexity, Game Theoretical Equilibrium, Maximum Entropy, Robustness, Exponential Families, Bregman Generator.
PACS: 65.40.Gr

THE PRINCIPLE OF GAME THEORETICAL EQUILIBRIUM

Consider a discrete *alphabet* $\mathbb{A}$ and *probability distributions* $P,Q,\cdots$ over $\mathbb{A}$. The set of all such distributions is denoted $M_+^1(\mathbb{A})$. A distribution is identified by its point probabilities: $P=(p_i)_{i\in\mathbb{A}}$. A *measure of complexity* is a map which to each pair (P,Q) of distributions assigns a value $\Phi(P,Q)\in[0,\infty]$ such that, for each $P\in M_+^1(\mathbb{A})$, the minimal value of $\Phi(P,Q)$ with $Q\in M_+^1(\mathbb{A})$ is assumed on the diagonal, i.e. for $Q=P$ and nowhere else unless $\Phi(P,P)=\infty$.

A *preparation* is any non-empty subset $\mathscr{P}\subseteq M_+^1(\mathbb{A})$. When $\mathscr{P}$ is fixed, a *consistent distribution* is a distribution in $\mathscr{P}$. The *game* $\gamma=\gamma(\Phi,\mathscr{P})$ has Φ as objective function and is the two-person zero-sum game between *Player I* ("Nature"), who can choose a strategy $P\in\mathscr{P}$, and *Player II* ("the Physicist") who can choose any strategy $Q\in M_+^1(\mathbb{A})$. Player I is a maximizer, Player II a minimizer. Thus val_I defined by $\mathrm{val}_I=\sup_{P\in\mathscr{P}}\inf_Q\Phi(P,Q)$ is the *Player I-value* of the game and, similarly, val_{II} defined by $\mathrm{val}_{II}=\inf_Q\sup_{P\in\mathscr{P}}\Phi(P,Q)$ is the *Player II-value* of the game. Here and below, a variable denoted by Q is understood to vary over all of $M_+^1(\mathbb{A})$.

An *optimal Player I-strategy* is a $P\in\mathscr{P}$ such that $\mathrm{val}_I=\inf_Q\Phi(P,Q)$ and an *optimal Player II-strategy* is a $Q\in M_+^1(\mathbb{A})$ such that $\mathrm{val}_{II}=\sup_{P\in\mathscr{P}}\Phi(P,Q)$. By the general *minimax inequality*, $\mathrm{val}_I\le\mathrm{val}_{II}$. The game is in *equilibrium* if $\mathrm{val}_I=\mathrm{val}_{II}<\infty$.

For further information about the game introduced, see [4]. The attempt to locate optimal strategies for the players and to establish equilibrium for suitable preparations is taken as a basic principle of statistical physics, the *principle of game theoretical equilibrium* (GTE).

We introduce Φ-*entropy* of P as minimal complexity, i.e. as $\mathrm{H}(P)=\inf_Q\Phi(P,Q)$. By assumption, $\mathrm{H}(P)=\Phi(P,P)$, thus, $\mathrm{val}_I=\sup_{P\in\mathscr{P}}\mathrm{H}(P)$, which is the *maximum entropy value*, also denoted $\mathrm{MaxEnt}=\mathrm{MaxEnt}(\Phi,\mathscr{P})$. So $\mathrm{val}_I=\mathrm{MaxEnt}$ and we realize that the GTE-principle leads directly to Jaynes *maximum entropy principly*, cf. [5].

Classical Boltzmann-Gibbs-Shannon entropy (BGS-entropy) is obtained as minimal complexity with respect to the measure $(P,Q) \curvearrowright \sum p_i \ln \frac{1}{q_i}$ which has a clear and convincing interpretation related to coding. Our results go some way to establish reasonable interpretations also for more general measures of complexity. Regarding the origin of the the above measure of complexity, under the name of *inaccuracy*, see Kerridge [6].

As we have seen, entropy is generated by complexity. So is *divergence (cross entropy, relative entropy* or *redundancy)*, defined as actual minus minimal complexity: $\mathrm{D}(P,Q) = \Phi(P,Q) - \mathrm{H}(P)$ when $\mathrm{H}(P) < \infty$. In any case, the *linking identity* $\Phi(P,Q) = \mathrm{H}(P) + \mathrm{D}(P,Q)$ holds and $\mathrm{D}(P,Q) \geq 0$ with equality if and only if $P = Q$ (for the measures of complexity we shall consider, it will be clear how to define $\mathrm{D}(P,Q)$ when $\mathrm{H}(P) = \infty$).

ROBUSTNESS, EXPONENTIAL FAMILIES

A Player II-strategy Q is *robust* if, for some constant $h < \infty$, the *level of robustness*, $\Phi(P,Q) = h$ for all consistent distributions P. The set $\mathscr{E} = \mathscr{E}(\Phi, \mathscr{P})$ of all robust Player II-strategies is the *exponential family* associated with $\gamma(\Phi, \mathscr{P})$. If a family $\mathscr{N}$ of preparations is considered, the *exponential family* $\mathscr{E}(\Phi, \mathscr{N})$ associated with $\mathscr{N}$ is the set of distributions which are robust for all preparations $\mathscr{P} \in \mathscr{N}$.

The following general and simple observation will play a key role in the sequal:

Theorem 1 (robustness lemma). *Let the measure of complexity Φ and the preparation $\mathscr{P}$ be given. Assume that the distribution Q^* is robust ($Q^* \in \mathscr{E}(\Phi, \mathscr{P})$) and consistent ($Q^* \in \mathscr{P}$). Then $\gamma(\Phi, \mathscr{P})$ is in equilibrium and has Q^* as the unique MaxEnt-distribution as well as the unique optimal strategy for Player II.*

Proof. Though known from e.g. [4] we present a direct proof.

Let h be the level of robustness. Then $\Phi(Q^*, Q^*) = h$ and, for $P \in \mathscr{P}$ with $P \neq Q^*$, $\mathrm{H}(P) = \Phi(P,P) < \Phi(P,Q^*) = h$. Thus Q^* is the unique MaxEnt-distribution. For any $Q \neq Q^*$, $\sup_{P \in \mathscr{P}} \Phi(P,Q) \geq \Phi(Q^*,Q) > \Phi(Q^*,Q^*) = h = \sup_{P \in \mathscr{P}} \Phi(P,Q^*)$ and equilibrium as well as unique optimality of Q^* for Player II follows. □

The result connects the exponential family $\mathscr{E}$ with the preparation $\mathscr{P}$. Indeed, if $\mathscr{E}$ and $\mathscr{P}$ intersect, they only intersect in one distribution which then is the optimal strategy for both players and, furthermore, the game considered is in equilibrium.

COMPLEXITY AND LINEAR CONSTRAINTS

We shall apply the principle of GTE – via the robustness lemma – to a wide class of complexity functions and associated notions of entropy, always having one and the same type of preparations in mind, viz. those given by *linear constraints*. They are the most important preparations for statistical physics and other applications, cf. e.g. Kapur [7].

From now on, we consider a fixed finite set $f = (f_\nu)_{1 \leq \nu \leq k}$ of real-valued functions defined on $\mathbb{A}$. The associated *family of natural preparations*, denoted $\mathscr{N}$, consists of all non-empty sets $\mathscr{P}_a$ which are defined as follows, denoting by $\langle \cdot, P \rangle$ mean value w.r.t. P:

$$\mathscr{P}_a = \{P \in M_+^1(\mathbb{A}) | \langle f_\nu, P \rangle = a_\nu \text{ for } 1 \leq \nu \leq k\}. \tag{1}$$

Here $a = (a_\nu)_{1\le\nu\le k} \in \mathbb{R}^k$. We assume that no non-trivial linear combination of the f_ν's reduces to a constant function. Clearly, $\mathscr{E}(\Phi,\mathscr{N})$, the *natural exponential family*, consists of those distributions which are robust for all natural preparations.

We shall select special measures of complexity adapted to a study of the natural preparations and constructed with the aim to simplify the search for distributions in $\mathscr{E}(\Phi,\mathscr{N})$. To accomplish this, we consider measures of complexity of the form

$$\Phi(P,Q) = \xi_Q\Big(\langle\overline{\kappa}(Q),P\rangle\Big) \tag{2}$$

where, for each $Q \in M_+^1(\mathbb{A})$, ξ_Q is a real function and $\overline{\kappa}$ maps $Q \in M_+^1(\mathbb{A})$ into a function defined on $\mathbb{A}$. We insist that $\langle\overline{\kappa}(Q),P\rangle$ can be obtained by summation based on a function $\kappa : [0,1] \to [0,\infty]$, the *coding function*, via the formula

$$\langle\overline{\kappa}(Q),P\rangle = \sum_{i\in\mathbb{A}} p_i\kappa(q_i)\,. \tag{3}$$

This corresponds to the requirement $(\overline{\kappa}(Q))(i) = \kappa(q_i)\,;\, i \in \mathbb{A}$.

Regarding $\xi_Q : [0,\infty] \to [0,\infty]$ and $\kappa : [0,1] \to [0,\infty]$, we assume that the ξ_Q's are increasing and concave, that κ is decreasing and convex, that $\kappa(1) = 0$, that κ is continuous at 0 (not just at $]0,1]$) and, finally, that Φ defined by (2) is a genuine measure of complexity. The last requirement will be trivially fulfilled in the concrete cases we shall consider. The inverse function $\kappa^{-1} : [0,\kappa(0)] \to [0,1]$ will play a significant role. We note that this function is continuous, decreasing and convex, as is κ (simple geometric proof).

For the *classical example*, ξ_Q is the identity map and κ the function $q \curvearrowright \ln\frac{1}{q}$. Then κ^{-1} is the restriction of $x \curvearrowright \exp(-x)$ to $[0,\infty]$. Entropy generated by this measure of complexity is standard BGS-entropy.

For the general situation, we note that any Q for which $\overline{\kappa}(Q)$ is a linear combination of the constant function 1 and the given functions $f_1,\cdots,f_k$, i.e. of the form

$$\overline{\kappa}(Q) = \lambda_0 + \lambda_1\cdot f_1 + \cdots \lambda_k\cdot f_k = \lambda_0 + \lambda\cdot f \tag{4}$$

for certain constants λ_0 and $\lambda = (\lambda_1,\cdots,\lambda_k)$, is a member of $\mathscr{E}(\Phi,\mathscr{N})$. Motivated by this observation, we fix real constants $\lambda = (\lambda_1,\cdots,\lambda_k)$ and ask if there exists a real constant λ_0 and a distribution $Q = (q_i)_{i\in\mathbb{A}}$ such that (4) holds.

For abbreviation, put $L_i = \lambda\cdot f(i)$. Then (4) amounts to $q_i = \kappa^{-1}(\lambda_0 + L_i)$ for $i \in \mathbb{A}$. As κ^{-1} is defined on $[0,\kappa(0)]$, we must have $0 \le \lambda_0 + L_i \le \kappa(0)$ for each i. Therefore, the L_i must be bounded below. Furthermore, from $\sum_i q_i = 1$, we conclude that, for each $K < \kappa(0)$, there can only be finitely many $i \in \mathbb{A}$ with $L_i \le K$. Thus we may order the L_i: $L_{i_1} \le L_{i_2} \le \cdots$, with this sequence breaking off and having a largest element if $\mathbb{A}$ is finite and with $L_{i_n} \to \kappa(0)$ if $\mathbb{A}$ is infinite. Put $L_* = L_{i_1}$ and $L^* = \sup_{i\in\mathbb{A}} L_i$ ($= \kappa(0)$ if $\mathbb{A}$ is infinite). We realize that we must require that $L^* - L_* \le \kappa(0)$ and, assuming this holds, the set of possible constants λ_0 is the set $[-L_*,\infty[$ in case $\kappa(0) = \infty$ and the set $[-L_*,\kappa(0) - L^*]$ if $\kappa(0) < \infty$. Consider the function f defined by $f(x) = \sum_{i\in\mathbb{A}} \kappa^{-1}(x + L_i)$ with x's ranging over the possible values of λ_0. What we search for is a value of λ_0, necessarily unique, such that $f(\lambda_0) = 1$.

Clearly, $f(-L_*) \geq 1$. By standard techniques, we see that f is continuous from the right and if $f(x_0) < \infty$ for some value of x_0, then f is continuous at all $x > x_0$. Furthermore, if $x_n \to \kappa(0)$ and if $f(x_n) < \infty$ for all n, then $f(x_n) \to 0$ as $n \to \infty$.

Our analysis shows that f can have at most one point of discontinuity, viz. where it passes from the value ∞ to finite values. Such a discontinuity "normally" does not occur. Also other anomalies are "normally" excluded. For instance, one may easily construct examples such that f is constantly equal to ∞ but such values are also excluded as they are of no practical interest. Thus we maintain that "normally" the function f assumes finite values larger than 1 as well as values less than 1 and hence the existence of a value λ_0 with $f(\lambda_0) = 1$ is assured by continuity.

Summarizing, we can now formulate the main result:

Theorem 2 (MaxEnt calculus). *Let $\lambda = (\lambda_1, \cdots, \lambda_k)$ be given real constants. Then, under "normal" circumstances (cf. the discussion above), the equation*

$$\sum_{i \in \mathbb{A}} \kappa^{-1}\Big(\lambda_0 + \lambda \cdot f(i)\Big) = 1 \tag{5}$$

has a solution, necessarily unique, and $Q = (q_i)_{i \in \mathbb{A}}$ given by

$$q_i = \kappa^{-1}\Big(\lambda_0 + \lambda \cdot f(i)\Big) \text{ for } i \in \mathbb{A} \tag{6}$$

satisfies (4) *and hence belongs to the exponential family $\mathscr{E}(\Phi, \mathscr{N})$. This distribution is the MaxEnt-distribution for $\mathscr{P}_a$ with $a = (a_1, \cdots, a_k)$ given by*

$$a_\nu = \sum_{i \in \mathbb{A}} q_i f_\nu(i) \text{ for } \nu = 1, \cdots, k \tag{7}$$

and, for this value of a, $\mathrm{MaxEnt}(\Phi, \mathscr{P}_a) = \xi_Q(\lambda_0 + \lambda \cdot a)$.

The theorem replaces and expands the standard recipe for MaxEnt-calculations. The main difference is a focus on λ_0 via (5) rather than on the classical partition function. In the final section we present a more thorough discussion of the significance of the result.

Before continuing, we shall limit the type of complexity functions studied by reducing the number of parameters needed for their definition. Instead of the many functional parameters appearing in (2), we now suggest a setting with only two functional parameters, one function ξ, called the *corrector*, to account for all the functions ξ_Q via the formula $\xi_Q(x) = x + \sum_{i \in \mathbb{A}} \xi(q_i)$ and then the already introduced coding function κ. In other words, we point to complexity functions of the form

$$\Phi(P, Q) = \sum_{i \in \mathbb{A}} p_i \kappa(q_i) + \sum_{i \in \mathbb{A}} \xi(q_i). \tag{8}$$

The functions κ and ξ are uniquely determined from Φ. The two terms in (8) are called, respectively the *coding part* and the *correction*. For the classical example, the coding part is $\sum_i p_i \ln \frac{1}{q}$ and the correction vanishes.

COMPLEXITY À LA BREGMAN

We shall now generate a $(\Phi, \mathrm{H}, \mathrm{D})$-triple from a simple starting point. The method follows the idea of *Bregman divergences* and is referred to as *Bregman generation*. Another method, *Csiszár generation*, was suggested in [4]. In our view, Bregman generation is by far the most important one for the needs of statistical physics.

Given is a *Bregman generator* by which we shall understand a strictly concave and smooth real function h defined on $[0,1]$ with $\mathrm{h}(0) = \mathrm{h}(1) = 0$ and $\mathrm{h}'(1) = -1$. We take "smoothness" to mean that h has an analytic extension to $[0, \infty[$. Though less will do for most investigations, the stronger requirement allows one to consider also the *dual function* $\tilde{\mathrm{h}}$ defined by

$$\tilde{\mathrm{h}}(x) = x\,\mathrm{h}\left(\frac{1}{x}\right). \tag{9}$$

This function is well-defined and real-valued in $]0, \infty[$. As a final technical assumption, we assume that the function can be extended by continuity to $[0, \infty]$, allowing for infinite values at the endpoints. A specific value $\mathrm{h}(p)$ is interpreted as the *complexity* of an event which is known to occur with probability p.

From h we generate two functions, $\phi = \phi(p,q)$, and $\mathrm{d} = \mathrm{d}(p,q)$:

$$\phi(p,q) = \mathrm{h}(q) + (p-q)\,\mathrm{h}'(q), \tag{10}$$

$$\mathrm{d}(p,q) = \mathrm{h}(q) - \mathrm{h}(p) + (p-q)\,\mathrm{h}'(q). \tag{11}$$

A specific value $\phi(p,q)$ is interpreted as the *complexity* of an event which is believed to occur with probability q but actually occurs with probability p. This is consistent with the previous interpretation as $\phi(p,p) = \mathrm{h}(p)$. The function d simply measures the difference (*divergence*) between estimated and true value. We also note that $\phi(p,q)$ and $\mathrm{d}(p,q)$ may assume the value $+\infty$. This happens if and only if both $p > q = 0$ and $\mathrm{h}'(0) = \infty$ hold.

Consider the *internal* functions, $\Phi = \Phi_{\mathrm{h}}$, $\mathrm{H} = \mathrm{H}_{\mathrm{h}}$ and $\mathrm{D} = \mathrm{D}_{\mathrm{h}}$ generated by ϕ, h and d. By this we mean that:

$$\Phi(P,Q) = \sum_{i \in \mathbb{A}} \phi(p_i, q_i)\,,\ \mathrm{H}(P) = \sum_{i \in \mathbb{A}} \mathrm{h}(p_i)\,,\ \mathrm{D}(P,Q) = \sum_{i \in \mathbb{A}} \mathrm{d}(p_i, q_i)\,. \tag{12}$$

We refer to ϕ, h and d as the *partial* functions, respectively partial *complexity, entropy* and *divergence*. They satisfy a partial version of the linking identity:

$$\phi(p,q) = \mathrm{h}(p) + \mathrm{d}(p,q)\,. \tag{13}$$

Note that $\Phi = \Phi_{\mathrm{h}}$ is of the special form (8) with coding function $\kappa = \kappa_{\mathrm{h}}$ given by

$$\kappa(x) = \mathrm{h}'(x) + 1 \tag{14}$$

and corrector $\xi = \xi_{\mathrm{h}}$ given by $\xi(x) = \mathrm{h}(x) - x(\mathrm{h}'(x) + 1)$. Hence the Bregman generator is decomposed into two terms:

$$\mathrm{h}(x) = x\kappa(x) + \xi(x)\,. \tag{15}$$

As $\xi(0)=\xi(1)=0$ and $\xi'(x)=-x\mathrm{h}''(x)-1$ we find that $\xi\equiv 0$ if and only if we are in the *classical case* $h(x)=x\ln(1/x)$. We also see that $\xi(x)\geq -x$ in $[0,1]$, hence the correction related to any distribution Q is bounded below by -1. The dual function $\tilde{\mathrm{h}}$ appears also to be of significance. In particular, $\xi(x)=\tilde{\mathrm{h}}'(1/x)-x$, hence

$$\Phi(P,Q)=\sum_{i\in\mathbb{A}} p_i\mathrm{h}'(q_i)+\sum_{i\in\mathbb{A}}\tilde{\mathrm{h}}'(\frac{1}{q_i})\,. \tag{16}$$

The first term in (16) is the coding part minus 1, the second term the correction plus 1. Partial complexity is given by $\phi(p,q)=p\mathrm{h}'(q)+\tilde{\mathrm{h}}'(1/q)$.

GENERATORS VIA DEFORMED LOGARITHMS

We turn to a concrete two-parameter family $(\mathrm{h}_{\alpha,\beta})$ of Bregman generators defined via *deformed logarithms* (taken in this form from [10]) and given by

$$\ln_{\alpha,\beta}x=\begin{cases}\frac{x^\beta-x^\alpha}{\beta-\alpha}\text{ for }\alpha\neq\beta\\ x^\alpha\ln x\text{ for }\alpha=\beta\end{cases}\quad. \tag{17}$$

The associated Bregman generators are defined by

$$\mathrm{h}_{\alpha,\beta}(x)=x\ln_{\alpha,\beta}(1/x)\,. \tag{18}$$

Warning: We have chosen to model the definition after the expression $x\ln(1/x)$ rather than $-x\ln x$. The main reason is the more natural interpretation of the former expression, but also, the change appears to be more as preferred in the "Tsallis literature". The change is in contrast to the choice in [4]. Thus, compared to [4], one should make the transformation $(\alpha,\beta)\curvearrowright(-\beta,-\alpha)$. Note also the symmetry $\mathrm{h}_{\alpha,\beta}=\mathrm{h}_{\beta,\alpha}$.

From [4] we see (after transformation) that, in order to obtain a genuine Bregman generator, the following restrictions apply to α and β: *Either* $0\leq\alpha<1$ and $\beta\leq 0$ *or else* $\alpha\leq 0$ and $0\leq\beta<1$.

The partial complexity function and the coding function are given by:

$$\phi_{\alpha,\beta}(x,y)=\frac{1}{\beta-\alpha}\Big(-(1-\alpha)xy^{-\alpha}+(1-\beta)xy^{-\beta}-\alpha y^{1-\alpha}+\beta y^{1-\beta}\Big)\,, \tag{19}$$

$$\kappa_{\alpha,\beta}(x)=1-\frac{1}{\beta-\alpha}\Big((1-\alpha)x^{-\alpha}-(1-\beta)x^{-\beta}\Big)\,. \tag{20}$$

Note that $\kappa(0)=\infty$ except if either $\alpha=0$ or $\beta=0$ (then $\kappa(0)=(\alpha+\beta-1)/(\alpha+\beta)$).

The important inverse functions κ^{-1} are defined on $[0,\kappa(0)]$. They can only be calculated in closed form in special cases. We point to the *Tsallis case* which corresponds to $\alpha<1$, $\beta=0$. The *Tsallis parameter*, traditionally denoted by q, is then given by $q=1-\alpha$. For the origin to this family within the physics literature, see Tsallis, [11]. Let us put $\kappa_{\alpha,0}=\kappa_q$ (as above with $q=1-\alpha$). Then, for $q\neq 1$,

$$\kappa_q^{-1}(x)=\Big(1+\frac{1-q}{q}x\Big)^{\frac{1}{q-1}}\text{ for }0\leq x\leq\kappa_q(0) \tag{21}$$

and one can insert (21) into (5). The kind of sums obtained will, typically, have to be calculated numerically. An exception is the case $q = 2$. We leave it to the reader to work out the pleasent details of our calculus in this case (take $\mathbb{A}$ to be finite).

Another case where $\kappa^{-1}_{\alpha,\beta}$ can be calculated in closed form is the *Kaniadakis family* which corresponds to $\alpha = -\beta$, cf. Kaniadakis [12]. We shall not go into that here.

DISCUSSION

Some features of the main result. Theorem 2 provides a theoretical framework for MaxEnt calculations for natural preparations given by linear constraints and pertaining to a wide range of different entropy measures. Among special features as compared with the standard approach we mention the following:

The basis for the result is the game theoretical approach which necessitates a focus on possibly unfamiliar aspects and quantities, notably a focus on a notion of complexity, intended to reflect the interplay between the physicist and the system he is studying. This aspect could have been hidden, but the underlying principle – the principle of *Game Theoretical Equilibrium* – is in itself promoted as a major issue. Indeed, it is suggested that this principle is of a basic nature, applicable to several scientific investigations, and that, for the area of statistical physics, it is more fundamental than Jaynes Maximum Entropy Principle. The principle originated with Pfaffelhuber [13] and, independently, the author (with [14] the first publication in English). Among further studies, we mention the joint work [15] with Harremoës.

Another feature is the puzzling fact that optimization has been achieved "miraculously" without recourse to Lagrange multipliers. Many will find it difficult to accept that for the problem studied, an approach which is better – simpler and more illuminating – than the well proven technique involving the popular multipliers exists. Within the mathematical literature, this special feature goes back at least to Csiszár, cf. [23].

Finally, we note that the MaxEnt calculus outlined here has no mention of partition functions. The calculus goes a good deal beyond traditional settings based on classical BGS-entropy. This has resulted in a focus on λ_0 which corresponds to the logarithm of the partition function in the classical case (so, for the classical case, we can write $\lambda_0 = \ln Z(\lambda)$ where $Z(\lambda) = \sum \exp(-\lambda \cdot f(i))$). It is well known that $\ln Z$ is a key quantity to work with, thus this feature should be no great surprise. But it is interesting that our approach leads directly to this quantity. As the partition function has no place for the general case covered by Theorem 2, this is of course also forced in some sense.

Exponential families. Whereas the concept of partition function does not survive the extension to general entropy- and complexity measures, the notion of *exponential families* does. It even appears to be *the* central concept behind the approach taken, cf. Theorem 1. However, extensions of this concept are needed (see below).

Comparing with the classical approach. The simplifications in the classical case result from the factorization property of κ^{-1}, an exponential function in that case. Apart from this, the calculations for a general complexity function appear to be of much the same nature as for the classical case. Indeed, given $\lambda = (\lambda_1, \cdots, \lambda_k)$ one determines λ_0 from (5) and then, via (6), (7) leads to the relevant averages $a = (a_1, \cdots, a_k)$. If you aim for a specific set of averages, there seems to be no way, neither in the classical case nor in the

general setting, other than application of numerical optimization procedures to choose just that set of parameters λ which leads to the appropriate set of constrained values. This discussion then tells us that apart from the simplifications possible in handling (5), the general calculus suggested is no more complicated in practise than what you are used to from classical studies.

Thermodynamic calculus. The difficulties, indeed impossibilities, involved in finding solutions to MaxEnt problems in closed form for other than the simplest problems constitute part of the motivation to create a thermodynamic calculus, studying variation as functions of various parameters of significance to the physicist or chemist. In this way one hopes to develop useful approximate solutions or to discover interesting trends in the thermodynamics as response to changes of relevant parameters. The differential calculus needed for such endeavours appears to be applicable also to the general setting of Theorem 2 with its precise equations to look closer into. Studies of this kind are not taken up here.

Natural expansions, optimal opdating based on a prior. There are many further possibilities for theoretical investigations based on measures of complexity of the form here studied. Assumptions related to the form (2) allows one to derive several results other than Theorem 2: Uniqueness of Q determined from λ, convexity of the set of λ's for which Q can be found, convexity of the function $\lambda \curvearrowright \lambda_0 = \lambda_0(\lambda)$ (this corresponds in the classical case to log-convexity of the partition function), existence of equilibria for the models in the natural family and, as a consequence, concavity of the map $a \curvearrowright \mathrm{MaxEnt}(\Phi, \mathscr{P}_a)$.

We comment that whereas measures of complexity of the special form (8) are rather simple and quite a rich family, the more elaborate form given by (2) is also of importance – especially, it allows the consideration of *Rényi entropies* and related quantities.

A special expansion of the concept of robustness which allows identification of MaxEnt-distributions for which some of the point probabilities (the q_i of Theorem 2) are allowed to be 0 should also be mentioned. This concerns cases where $\lambda_0 + \lambda \cdot f(i) \geq \kappa(0)$ and is therefore only relevant when $\kappa(0) < \infty$. However, there are important cases where this is so, e.g. Tsallis-type quantities with $q > 1$. In such cases *inconsistent* inference is possible where a *feasible i* (one for which there exists $P \in \mathscr{P}_a$ with $p_i > 0$) is inferred under MaxEnt-based inference as an impossible event. This phenomenon is treated in part by Jaynes, cf. p.345 of [22]. Taking this into consideration, it appears possible to prove that any candidate to MaxEnt-distributions (or the more general *centers of attraction* of [15]) of preparations in a natural family of preparations, must be a member of the associated exponential family. For the classical case, where inconsistent inference is not possible, such a result was established in [15].

Consider now the problem of optimal *updating* based on a given *prior*. In fact, such problems can be handled in analogy with our analysis of MaxEnt problems. In particular, a result à la Theorem 2 holds which provides a calculus for optimal *posterior distributions* via a *minimum cross entropy principle* – the kind of results initiated by Kullback, cf. [24]. To indicate, if only briefly, that this requires no new techniques, consider a prior Q_0 and try to maximize the *updating gain* $\Psi_{Q_0}(P,Q) = \Phi(P,Q_0) - \Phi(P,Q)$. This situation can be analyzed by applying our game theoretical reasoning to $-\Psi_{Q_0}$ which is a genuine complexity measure. For this to work, the theory has to be extended slightly, allowing complexity measures that can take negative values.

Precise statements and proofs of results just indicated will be published elsewhere.

Origin of the two-parameter family. The two-parameter family of complexity-, entropy- and divergence measures, $(\Phi_{\alpha,\beta}, \mathrm{H}_{\alpha,\beta}, \mathrm{D}_{\alpha,\beta})$ has its origin in the mathematical literature, cf. Mittal [8] and Sharma and Taneja [9], and was studied later in the physical literature by Borges and Roditi, [10] who used the convenient concept of deformed logarithms.

Entropy should not stand alone. Let us illustrate this thesis by considering Tsallis entropy with Tsallis parameter q. There are infinitely many ways of obtaining this entropy measure as minimal complexity. Below we suggest three complexity measures which have this property:

$$\Phi^B(P,Q) = \frac{1}{q-1} + \sum \left(q_i^q - \frac{q}{q-1} p_i q_i^{q-1} \right) \tag{22}$$

$$\Phi^C(P,Q) = \frac{1}{1-q} \sum p_i^q (1 - q_i^{1-q}) \tag{23}$$

$$\Phi^R(P,Q) = \frac{1}{1-q} \left(\frac{\sum p_i^q}{\sum p_i^q q_i^{1-q}} - 1 \right). \tag{24}$$

As usual, sums are over $i \in \mathbb{A}$. The "B ", "C" and "R" stand for, respectively "Bregman " , "Csiszár" and "Rényi" . The complexity measure Φ^B is the one considered in the main text, Φ^C the one considered in [4] and Φ^R is closely related to the relevant complexity measure connected with Rényi entropy and divergence.

The measure Φ^B allows us – as we have seen – to study the natural preparations given by linear constraints, Φ^C allows us to develop a calculus much as Theorem 2, but aiming at maximizing entropy for preparations given by averaging with respect to the *q-associated measures* which are measures with point masses p_i^q and finally, Φ^R allows us to deal with preparations given by averages with respect to the *q-escort distributions* which are obtained by normalizing the q-associated measures. To realize that this is indeed so, you just have to note how P enters in the complexity measure considered. It can safely be argued that "distorted" averages as those indicated above related to Φ^C and Φ^R have no physical relevance and therefore, they are considered of less or no importance for the study of natural maximum entropy problems. Bregman generation is thus the method which stands back as the really significant method.

The importance of Bregman type quantities. The relevance for statistical physics of Bregman divergence was emphasized by Naudts [1], [2]. The work by Abe and Bagci [3] should also be mentioned, however, the present author does not agree with their conclusion that the use of escort distributions is essential. Anyhow, the proper matching of entropy measure with the type of constraints one wants to study is important. This issue is also addressed in Feng [20].

Originally, Bregman introduced the concept to meet needs of learning theory, cf. [21]. For more recent articles in this direction, see Murata et al., [19] and Sears [18].

Concerning extensions in another direction, to quantum statistical physics, note the recent study by Petz, [17] where Bregman divergences are carefully defined. Incorporation of game theoretical considerations may be a fruitful area of research to look into.

Interpretations. Any measure of entropy of importance to statistical physics should be motivated by sound reasons, including appropriate interpretations. It appears that

Bregman generation in itself goes a way in this direction. In addition, the choice of terminology, especially regarding the frequent reference to "coding" , though not yet founded in precise procedures for observation or measurement, is indicative for what future research may bring, at least this is where speculations of the author goes.

One should recall that Kullback-Leibler divergence is related to free energy for classical preparations. This kind of interpretation when more general Bregman-type divergences are involved appears also to be sound, cf. the recent study by Bagci, [16]. Possibly, Crooks, [25], also points to issues to be integrated before a full picture is in place.

ACKNOWLEDGMENTS

Thanks are due to Jan Naudts for critical comments and suggestions and to Bjarne Andresen for pointing me to [25]. The work was supported by the Danish Natural Science Research Council.

REFERENCES

1. J. Naudts. *Rev. Math. Phys.*, 16(6):809–822, 2004.
2. J. Naudts. *J. Ineq. Pure and Appl. Math.*, 5(4):1–15, 2004.
3. S. Abe and G. B. Bagci. *Physical Review E*, 71:016139,1–5, 2005.
4. F. Topsøe. *Physica A*, 340/1-3:11–31, 2004.
5. E. T. Jaynes. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.
6. D. F. Kerridge. *J. Roy. Stat. Soc. B.*, 23:184–194, 1961.
7. J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley, New York, 1993. first edition 1989.
8. D. P. Mittal. *Metrika*, 22:35–45, 1975.
9. B. D. Sharma and I. J. Taneja. *Metrika*, 22:205–215, 1975.
10. E. P. Borges and I. Roditi. *Physics Letters A*, 246:399–402, 1998.
11. C. Tsallis. *J. Stat. Physics*, 52:479, 1988. See http://tsallis.cat.cbpf.br/biblio.htm for a comprehensive bibliography.
12. G. Kaniadakis. *Physical Review E*, 66:056125,1–17, 2002.
13. E. Pfaffelhuber. Minimax information gain and minimum discrimination principle. In I. Csiszár and P. Elias, editors, *Topics in Information Theory*, volume 16 of *Colloquia Mathematica Societatis János Bolyai*, pages 493–519. János Bolyai Mathematical Society and North-Holland, 1977.
14. F. Topsøe. *Kybernetika*, 15(1):8 – 27, 1979.
15. P. Harremoës and F. Topsøe. *Entropy*, 3(3):191–226, Sept. 2001.
16. G. B. Bagci. arXiv:cond-mat/0703008v1, March 2007.
17. D. Petz. *Acta Math. Hungar.*, 116:127–131, 2007.
18. T. D. Sears. From maxent to machine learning and back. In K. Knuth, editor, *Proceedings of the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2008. to appear.
19. T. Kanamori N. Murata, T. Takenouchi and S. Eguchi. *Neural Computation*, 16(7):1437–1481, 2004.
20. X. Feng. arXiv:cond-mat.stat-mech/0705.1332v4, May 2007.
21. L. M. Bregman. *USSR Comp. Math. and Math. Phys.*, 7:200–217, 1967.
22. E. T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, 2003.
23. I. Csiszár. *Ann. Probab.*, 3:146–158, 1975.
24. S. Kullback. *Informaton Theory and Statistics*. Wiley, New York, 1959.
25. G. E. Crooks. *Physical Review Letters*, 99:100602, 1–4, 2007.