International Journal of Theoretical Physics, Vol. 37, No. 1, 1998

On Informational Divergences for General Statistical Theories

Stephan Zanzinger[1]

Received July 4, 1997

We describe a way to transfer informational divergences into the nonclassical regime, and state their basic properties. This should be seen as a first step toward a nonclassical estimation theory. Our procedure mimics the approach of V. Cantoni in defining a generalized transition probability and only needs a reasonable concept of observables (POV measures).

Informational divergences, in classical statistics also called *f*-divergences, are introduced in order to compare probability measures—the states of classical theory. In fact, these informational divergences measure the amount of information contained in the states and they are therefore a generalization of the relative entropy, which is called in information theory also *relative information.* The obtained variety of functionals contains the relative entropy (relative information), the Rényi entropies, and the Bures distance, which is associated to the generalized transition probability introduced by Cantoni (1975). The aim of the present paper is to transfer these functionals into the quantum world. The plan of the paper is as follows. First, we refer the definition of *f*-divergences of classical statistics, touch upon some of their properties, and give the basic examples as they appear in statistics. Then we switch to general statistical theories, where we introduce the basic concepts and state some further properties of our *f*-divergences. Finally, we show how statistical divergences are introduced in quantum probability and compare the ansatz found there with our ideas. We omit the proofs and leave various mathematical objects and notation unexplained; for the classical part we therefore recommend Liese and Vajda, (1985) and Heyer (1982), for issues

[1] Institut für Theoretische Physik, Universität Tübingen, D-72076 Tübingen, Germany.

0020-7748/98/0100-035 7$15.00/0

of quantum probability our basic reference is Ohya and Petz (1993); details will be elaborated in a forthcoming publication (see also Zanzinger, 1995).

The basic ingredient for the definition of *f*-divergences is a *convex function*[2] $f\colon \mathbb{R}_0^+ \to \mathbb{R} \cup \{\infty\}$ as a bias for the comparison. By means of the auxiliary function

$$\tilde{f}(u, v) := \begin{cases} 0 & \text{if } u = 0 \text{ and } v = 0 \\ u\dfrac{f(\infty)}{\infty} & \text{if } u \neq 0 \text{ and } v = 0 \\ vf\left(\dfrac{u}{v}\right) & \text{otherwise} \end{cases} \tag{1}$$

we define the *f*-divergence between two positive measures $\mu, \nu \in \mathcal{M}^+(\mathcal{X}, \Sigma)$ (probability distributions) on a measurable space $(\mathcal{X}, \Sigma)$:

$$f(\mu\|\nu) = \int_\chi \tilde{f}\left(\frac{d\mu}{d\sigma}, \frac{d\nu}{d\sigma}\right) d\sigma = \int_\chi f\left(\frac{d\mu_1}{d\nu}\right) d\nu + \mu_2(\chi)\frac{f(\infty)}{\infty} \tag{2}$$

where we decompose $\mu = \mu_1 \oplus \mu_2$ by means of the Lebesgue decomposition ($\mu_1 << \nu$ and $\mu_2 \perp \nu$). Here σ is an arbitrary positive measure to which μ and ν are absolutely continuous.

Theorem 1 (Range of Values). Let *f* be strictly convex. The following holds for probability measures μ and ν:

(i) $f(\mu\|\nu) \in [f(1), f(0) + f(\infty)/\infty]$.
(ii) $\mu = \nu \Leftrightarrow f(\mu\|\nu) = f(1)$.
(iii) $\mu \perp \nu \Rightarrow f(\mu\|\nu) = f(0) + f(\infty)/\infty$.
(iv) For $f(0) + f(\infty)/\infty < \infty$ the reverse implication in (iii) holds as well.

We see that an *f*-divergence is a good indicator for equality as well as for disjointness of states. At this point we give some examples which illustrate the unifying character of the concept of *f*-divergences. The most important representative, the *relative entropy*, is obtained by choosing $f(u) = -\ln(u) + u - 1$. In this case one gets

$$f(\mu\|\nu) =: S(\mu, \nu) = \begin{cases} -\displaystyle\int_\chi \ln\left(\frac{d\mu}{d\nu}\right) d\nu & \text{if } \mu << \nu \\ +\infty & \text{otherwise} \end{cases}$$

If we choose the (concave!) function $f(u) = H_\alpha(u) := u^\alpha$, we get the *Hellinger integral of order* $\alpha \in]0, 1[$:

$$f(\mu\|\nu) = H_\alpha(\mu\|\nu) = \int_\chi \left(\frac{d\nu}{d\sigma}\right)^\alpha \left(\frac{d\mu}{d\sigma}\right)^{1-\alpha} d\sigma$$

[2] We could as well choose concave functions.

The distinguished value $\alpha = \frac{1}{2}$ yields the *affinity* and leads to the *generalized transition probability* (Cantoni, 1975) in the quantum world. With the use of the Hellinger integrals the Rényi entropies of order α are defined by

$$S_\alpha(\mu, \nu) = \frac{1}{\alpha(\alpha - 1)} \ln H_\alpha(\mu \| \nu)$$

This (continuous) one-parameter family of functionals emerges, as the relative entropy, from an axiomatic codification of (relative) information (Rényi, 1961) and it holds that $\lim_{\alpha \to 0} S_\alpha(\mu, \nu) = S(\mu, \nu)$. This emphasizes the entropic character of the Hellinger integrals and therefore of the generalized transition probability (Gudder *et al.*, 1979).

In a general statistical theory the set of states is taken as primitive concept. It is modeled by a convex set $\mathcal{K}$ which is the base of a *base normed space* $\mathcal{V} = \mathcal{V}^+ - \mathcal{V}^+$ with positive cone $\mathcal{V}^+ = \cup_{\lambda \in \mathbb{R}^+} \lambda \mathcal{K}$ (Alfsen and Shultz, 1976). The dual space $\mathcal{V}^* = \mathcal{A}$—relevant for the definition of observables—equipped with the dual order and norm is an *order unit space.* The *order unit* e coincides on $\mathcal{V}^+$ with the norm. Elements of $[0, e] \subseteq \mathcal{A}$ are called *effects* and the extremal points $u \in \mathcal{U} = \partial_e[0, e]$—the generalizations of the projections of quantum mechanics—are called *decision effects.* As mentioned in the introduction, the concept of observables links general quantum theories to classical statistics and is the basic tool for the so-called *minimal statistical interpretation* (Busch *et al.*, 1991) of quantum mechanics. We stick here to the definition of observables as positive operator valued measures (POVM), also called *unsharp* observables:

Definition 2 (Observable). Let $(\mathcal{X}, \Sigma)$ be a measurable space. An $((\mathcal{X}, \Sigma)-)$ *observable is a map* A: $\Sigma \mapsto [0, e]$ *fulfilling*:

(i) $A(\mathcal{X}) = e$.

(ii) $A(\cup_{n \in \mathbb{N}} E_n) = \Sigma_{n \in \mathbb{N}} A(E_n)$ for each sequence $(E_n)_{n \in \mathbb{N}}$ of pairwise disjoint measurable sets.

Here the (infinite) sum is understood in the w^*-sense. An observable is called *sharp (PV)* if its range is contained in the set of decision effects.

Associated with the concept of observables is the following interpretation. An observable A represents a measurement with possible outcomes in the value space $\mathcal{X}$. If the system is in the state ω and an observable A is measured, then the number $\langle \omega; A(E) \rangle$ is the probability that the outcome of the measurement lies in the set $E \subseteq \mathcal{X}$. The basic idea that observables connect general statistical theories to classical statistics is made clear by the following formalization, which is immediate by the definition of an observable.

Proposition 3. Let A be an $(\mathcal{X}, \Sigma)$-observable. Then T_A: $\mathcal{K} \mapsto \mathcal{M}_1^+(\mathcal{X}, \Sigma)$ defined by $(T_A\omega)(\cdot) := \langle \omega; A(\cdot) \rangle$ is an affine map from the state space $\mathcal{K}$

into the space of all probability measures $\mathcal{M}_1^+(\mathcal{X}, \Sigma)$ on $(\mathcal{X}, \Sigma)$. The other way round, each such affine map yields an observable. Thus, the map $A \mapsto T_A$ is one-to-one.

Using this technique, it is easy to transfer concepts of classical statistics into the nonclassical world. This is done by first taking into account each observable separately and then doing some optimization procedure. To make this explicit for the informational divergences between two states ω and φ we choose an observable A and compare the associated measures by defining

$$f_A(\omega\|\varphi) := f(T_A\omega\|T_A\varphi) \tag{3}$$

We call f_A the f-divergence between ω and φ induced by A. If we have a distinguished family of observables $\mathbb{O}$,[3] then

$$f_{\mathbb{O}}(\omega\|\varphi) := \sup\{f(T_A\omega\|T_A\varphi) |\ A \in \mathbb{O}\} \tag{4}$$

gives the maximal value of discrimination between the states ω and φ if one considers measurements of the observables in $\mathbb{O}$. Therefore it is reasonable in the case that $\mathbb{O}$ is the set off *all* observables to call $f_{\mathbb{O}}(\omega\|\varphi)$ *the* f-divergence $f(\omega\|\varphi)$ between ω and φ. If $\mathbb{O}$ is the set of all *sharp* observables, the associated f-divergence is denoted by $f_{PV}(\omega\|\varphi)$.

Now, for $f(\cdot\|\cdot)$ the range of values (cf. Theorem 1) remains exactly the same as in the classical regime if one defines $\omega \perp \varphi$ if there is an effect a with $\langle\omega;\ a\rangle = 1$ and $\langle\varphi;\ a\rangle = 0$. Relevant for the entropic character of the f-divergence is the *monotony*, which in the classical context is usually formulated in terms of stochastic kernels, which are not immediately at hand in our context (Schindler, 1991).

Proposition 4. For an affine mapping (coarse graining) between state spaces $T: \mathcal{K} \mapsto \mathcal{K}'$, i.e., $T^*: \mathcal{A}' \mapsto \mathcal{A}$ is positive and unit preserving, it holds that $f(\omega\|\varphi) \geq f(T\omega\|T\varphi)$.

This property is connected with the concept of sufficiency. Here, we want to call a family of observables $\mathbb{O}$ (f-) *sufficient* for two states ω and φ if $f_{\mathbb{O}}(\omega\|\varphi) = f(\omega\|\varphi)$ holds. Note, that we have chosen a different concept of sufficiency than in *quantum probability*. There (Ohya and Petz, 1993) one defines *sufficiency* in terms of completely positive mappings in analogy to *Blackwell* sufficiency in classical statistics, which is a more global notion. In particular, it is independent of the convex function f and implies sufficiency in our sense for all f. Now we can state approximation theorems as statements about sufficient sets of observables. For instance, the set of observables with *finite spectrum* is sufficient for all states and all convex functions f. In the

[3] For concave f we use here the infimum.

operator-algebraic context, we are often in the situation that the relevant state space $\mathcal{K}$ is the state space of a quasilocal C*-algebra $\mathfrak{A}_0^{\|\cdot\|}$, $\mathfrak{A}_0 := \cup_{\Lambda\in\mathfrak{L}} \mathfrak{A}_\Lambda$. Here the observables with range in $\mathfrak{A}_0$ (with finite spectrum), i.e., the strictly local ones, are sufficient for all states (see also Kosaki, 1983).

What is known about the set of the PV-observables, the observables in the traditional sense? In the classical scenario [$\mathcal{K} = \mathcal{M}(\mathcal{X},\Sigma)_1^+$] the sharp observables are indeed sufficient, which also ensures the consistency of our generalization. In the general case the question remains open, but we can state the following partial result.

Proposition 5. The following statements are equivalent:

(i) The sharp observables are sufficient.

(ii) For any coarse graining T: $\mathcal{K} \mapsto \mathcal{K}'$ it holds that $f_{\mathrm{PV}}(\omega\|\varphi) \geq f_{\mathrm{PV}}(T\omega\|T\varphi)$.

(iii) $f_{\mathrm{PV}}(\omega\|\varphi) \geq f_{\mathrm{PV}}(T_\omega\|T_\varphi)$ for any discrete, classical coarse graining T: $\mathcal{K} \mapsto (l^1)_1^+$.

That is, whenever f_{PV} is a "good" statistical functional, that is, it has the right categorial properties, the set of sharp observables is *f*-sufficient. We should remark at this point that the mentioned Hellinger integrals, therefore the generalized transition probability, fulfill—at least if $\mathcal{K}$ is *spectral* (Alfsen and Shultz, 1976)—the equivalent conditions of the above proposition. We now discuss further properties of the *f*-divergences, which rely on some deeper structural insights, and therefore impose on our state space $\mathcal{K}$ to be *projective* in the sense of Alfsen and Shultz (1976). That is, we have a rich (orthomodular) set of decision effects and to each decision effect there corresponds a filtering transformation, a so-called P-projection.

Proposition 6 (Convexity):

$$f(\lambda\omega_1 + (1-\lambda)\omega_2\|\lambda\varphi_1 + (1-\lambda)\varphi_2) \leq \lambda f(\omega_1\|\varphi_1) + (1-\lambda) f(\omega_2\|\varphi_2)$$

for $\omega_1, \omega_2, \varphi_1, \varphi_2 \in \mathcal{K}$ and $\lambda \in [0, 1]$. We have *equality* if $\omega_1, \varphi_1 \perp \omega_2, \varphi_2$.

If we have *commuting states* χ, φ, i.e., there are some common orthogonal decompositions $\chi = \Sigma\lambda_i\omega_i$, $\varphi = \Sigma\kappa_i\omega_i$, $\Sigma\lambda_i = 1 = \Sigma\kappa_i$, with ω_1 pairwise orthogonal, we conclude from the above proposition

$$f(\chi\|\varphi) = \sum \tilde{f}(\lambda_i, \kappa_i) = \sum_{\kappa_i\neq 0} f\left(\frac{\lambda_i}{\kappa_i}\right)\kappa_i + \sum_{\{i|\kappa_i=0\}} \lambda_i \frac{f(\infty)}{\infty}$$

In this case the *f*-divergences are determined by the statistical weights of the orthogonal decomposition, i.e., they have a *classical* interpretation. A similar situation occurs if we have *superselection sectors*, i.e., if we have an orthogonal set of split faces C_i with $\oplus_{i\in\mathbb{N}} C_i = \mathcal{K}$. If we consider the positive

functionals ω_i, φ_i, the unique components of the states ω, φ in C_i, we get $f(\omega\|\varphi) = \Sigma_{i\in\mathbb{N}} f(\omega_i\|\varphi_i)$. This expression inherits from the decomposition into sectors only the classical f-divergence of the statistical weights, which are hidden in the nonnormalized states ω_i, φ_i ($\|\omega_i\| = \langle\omega; C_i\rangle$). This shows the absence of quantum correlations between different superselection sectors. At this point one can speculate about more complex decompositions, for instance, by means of a split face valued measure $\hat{\mu}$. Formally one expects a decomposition "$f(\omega\|\varphi) = \int_X f(\omega(x)\|\varphi(x))\ \mu(dx)$". We must remark at this point that one gets into measure-theoretic trouble if one proceeds along this line of thought. But at least in the case of direct integrals of von Neumann algebras, where a measure-theoretic apparatus is at hand, the above formula can be made rigorous (Gerisch *et al.*, 1996).

Finally we switch to operator-algebraic quantum mechanics, i.e., $\mathcal{K}$ is represented by $\mathcal{M}^+_{*,1}$, the normal states of a W*-algebra $\mathcal{M}$. Here the notions of so-called *quantum probability* are applied. We get another generalization of f-divergences, which we sketch briefly. Fundamental is the *standard representation* $\langle\mathcal{M}, \mathcal{H}, \mathcal{P}, J\rangle$ (Strătilaă 1981), which gives a one-to-one mapping $\mathcal{V}^+ \leftrightarrow \mathcal{P}$ between the positive functionals and the self-dual cone $\mathcal{P}$. The role of the Radon–Nikodym derivative of classical statistics—our f-divergence is some function of this object—is played by the relative modular operator $\Delta_{\Omega\Phi}$. Now, the *quasi-entropy* f_Δ—another generalization of f-divergence into the nonclassical context—is defined in Ohya and Petz (1993) and Araki (1976) for the relative entropy by

$$f_\Delta(\omega\|\varphi) := f_{\Delta_{\Omega,\Phi}}(\omega_\Omega\|\omega_\Phi) = \langle\Phi|f(\Delta_{\Omega,\Phi})\Phi\rangle + \langle\Omega|(1 - s^{\mathcal{M}}(\Phi))\Omega\rangle \frac{f(\infty)}{\infty} \tag{5}$$

This statistical distance has similar properties to f-divergences if f is *operator-convex*. In connection with our definition of informational divergences the following holds:

$$f(\omega\|\varphi) \leq f_\Delta(\omega\|\varphi)$$

This results from a different definition of observables in quantum probability. There one regards a subalgebra of $\mathcal{M}$ as an observable. Our *sharp* observables are the *commutative* subalgebras. Thus, in quantum probability an optimization as in equation (3) is over a greater set of "observables." But, if ω and φ commute, we have

$$f(\omega\|\varphi) = f_{\mathrm{PV}}(\omega\|\varphi) = f_\Delta(\omega\|\varphi) \tag{6}$$

Here equation (6) indicates for special functions f—e.g., $f(u) = u^\alpha$—that ω and φ commute.

REFERENCES

Alberti, P. M. (1983). A note on the transition probability over C^*-algebras, *Letters in Mathematical Physics,* **7**, 25–32.

Alfsen, E. M., and Shultz, F. W. (1976). Non-commutative spectral theory for affine function spaces on convex sets, *Memoirs of the American Mathematical Society*, Vol. **172**.

Araki, H. (1976). Relative entropy of states of von Neumann algebras, *Publications of the RIMS,* **11**, 809–833.

Busch, P., Lahti, P., and Mittelstaedt, P. (1991). *The Quantum Theory of Measurement*, Springer, Berlin.

Cantoni, V. (1975). Generalized "transition probability," *Communications in Mathematical Physics,* **44**, 125–128.

Gerisch, T., Rieckers, A., and Zanzinger, S. (1996). Operator algebraic transition probabilities in macroscopic quantum systems, preprint.

Gudder, S., Marchand, J.-P., and Wyss, W. (1979). Bures distance and relative entropy, *Journal of Mathematical Physics,* **20**(9), 1963–1966.

Heyer, H. (1982). *Theory of Statistical Experiments*, Springer, Berlin.

Kosaki, H. (1983). On the Bures distance and the Uhlmann transition probability of states on a von Neumann algebra, *Proceedings of the American Mathematical Society,* **89**, 285.

Liese, F., and Vajda, I. (1985). *Convex Statistical Distances*, Teubner.

Ohya, M., and Petz, D. (1993). *Quantum Entropy and its Use*, Springer, Berlin.

Rényi, A. (1961). On measures of entropy and information, in *Proceedings of the 1st Berkeley Symposium*, Berkeley, Vol. I, pp. 547–561.

Schindler, C. (1991). Representations of state space transformations by Markov kernels, *International Journal of Theoretical Physics,* **30**(11), 1409–1431.

Strătilaă, S. (1981). *Modular Theory in Operator Algebras*, Abacus Press, Turnbridge Wells.

Zanzinger, S. (1995). Verallgemeinerte Übergangswahrscheinlichkeiten und Quasientropien in der Vielteilchenphysik, Ph.D. thesis, Universität Tübingen.