J Stat Phys (2007) 128: 1189–1196
DOI 10.1007/s10955-007-9344-7

How Should the Distance of Probability Assignments Be Judged?

Sumiyoshi Abe · Bernhard Lesche · Jens Mund

Received: 24 July 2006 / Accepted: 9 May 2007 /
Published online: 16 June 2007

Abstract It is shown that the l_1-distance in the space of probability assignments on a finite set Ω provides a criterion to judge whether two assignments are too close to each other to be distinguished by a statistical test. The criterion is independent of the number of elements of Ω. Other notions of distance are also discussed.

1 Introduction

Lately several authors investigated whether information measures such as generalized entropies are stable with respect to small variations of the probability assignment [1–8]. This issue arose investigating the concept of mixing character, which is the union of all information theoretical entropy concepts [9]. Particularly the investigation of the question of stability of Rényi entropies introduced the stability concept into literature [10]. This concept comprises the following simple idea: If S is some entropy defined on a space of probability assignments and S varies typically by an amount S_{TYP}, then $S(W_0)$ and $S(W_1)$ should not differ by an amount of the order of S_{TYP} if the probability assignments W_0 and W_1 are so close to each other that they cannot be distinguished by any feasible statistical test. In order to be able to formulate mathematical criteria of stability it is useful to have some kind of distance defined in the space of probability assignments. This distance should be one that permits to judge whether two assignments can or cannot be distinguished by means of statistical tests. As the application of these concepts to thermodynamic systems involves probability assignments defined on sets of an enormous number of micro-states (typically larger than $2^{(10^{23})}$) the relation between distance and testability should be one that does not depend on the number of micro-states. In the present work we shall show that the appropriate notion of distance is the well known l_1-distance.

S. Abe
Institute of Physics, University of Tsukuba, Ibaraki 305-8571, Japan

B. Lesche (✉) · J. Mund
Departamento de Física, Universidade Federal de Juiz de Fora, Juiz de Fora, MG 360 36 900, Brazil
e-mail: lesche@fisica.ufjf.br

2 Statement of the Problem

Let W_0 and W_1 be two probability assignments defined on a finite set Ω of micro-states. Suppose one of them, but we do not know which one, describes a physical ensemble, i.e. a class of objects furnished by independent and equivalent procedures, defined by some experimental prescription (this concept of ensemble corresponds to the concept of state of a system [11]). How can one perform an experiment to decide which one is the correct assignment? Such an experiment would experimentally distinguish the two assignments.

One may measure properties of a number Z of members of the ensemble and compare their relative frequencies with the probabilities as predicted by W_0 or W_1. Depending on whether the relative frequencies found correspond more with the probabilities as calculated according to W_0 or W_1 one will decide in favor of W_0 or W_1. The decision may be more or less reliable. For instance if W_0 and W_1 have disjoint support ($\forall x \in \Omega$: $W_0(x) \neq 0 \Rightarrow W_1(x) = 0$) one can, with a single experiment ($Z = 1$) and without doubt, decide which of the assignments is the correct one, by asking the experimental question: "$\boldsymbol{x} \in$ support of W_0?".[1] In general, however, one is limited to uncertain results and the number of experimental questions Z will have to be large in order to obtain a reasonable reliability. If the assignments W_0 and W_1 differ by such a small amount that the necessary Z becomes so large that a practical realization of the experiments becomes impossible, the assignments are physically indistinguishable. Our task is to relate the necessary Z-values with a distance of the assignments W_0 and W_1. In order to do so let us first explain some basic notions of mathematical statistics [12].

A decision procedure described above is called a test of a simple null-hypothesis H_0 (W_0 is the correct assignment) against the simple alternative H_1 (W_1 is the correct assignment). If one rejects the hypothesis H_0, despite the fact that H_0 is correct, one calls the wrong decision a mistake of the first kind. If one accepts H_0 erroneously this is called a mistake of the second kind. Let α_T and β_T be the probabilities to commit a mistake of the first and second kind in a test T, respectively:

$$\alpha_T = P_{H_0}(H_0 \text{ rejected}), \tag{1}$$

$$\beta_T = P_{H_1}(H_0 \text{ accepted}). \tag{2}$$

In these expressions $P_Y(X)$ designates the probability of the event X according to the hypothesis Y.

There are testing procedures that use a fixed sample size Z and others, called sequential tests, where the sample size depends on the experimental outcomes. Furthermore Z may also depend on other stochastic processes. In this case the test is called a randomized one. In this work we shall treat only tests with a fixed sample size. The extension to randomized and sequential tests is straightforward and does not bring essential new aspects into play.

The individual micro-states $x \in \Omega$ are not always observed directly. This is especially so in the case of thermodynamic systems. There one observes values of functions f defined on Ω. If one finds an experimental value $\boldsymbol{f}$ in an interval $[a, b]$ one says the micro-state has been found in the inverse image of that interval; $\boldsymbol{x} \in f^{-1}([a, b])$. So what can be observed is whether the micro-state is an element of certain subsets of Ω. The outcome of Z repeated experiments can be thought of as a composed micro-state that consists of a Z-tuple of elements of Ω; $\boldsymbol{X} = (x_1, x_2, \ldots, x_Z)$ with $x_i \in \Omega$. That means $\boldsymbol{X}$ is an element of the Cartesian

[1] We shall write experimental outcomes with bold face letter to distinguish them from other elements $x \in \Omega$.

product $\Omega^Z = \Omega \times \Omega \times \cdots \times \Omega$. Again, what can be observed is whether $\boldsymbol{X}$ is an element of certain subsets of Ω^Z.

A non-randomized test of sample size Z is defined by a subset $C_T \subset \Omega^Z$, called the critical set, and the rule that the null hypothesis H_0 is rejected if one finds experimentally $\boldsymbol{X} \in C_T$ and it is accepted otherwise. This way we get for the probabilities of the error of first and second kind

$$\alpha_T = P_{H_0}(\boldsymbol{X} \in C_T), \tag{3}$$

$$\beta_T = P_{H_1}(\boldsymbol{X} \notin C_T). \tag{4}$$

Fore instance in the case of the example of assignments W_0, W_1 with disjoint support one can choose $Z = 1$, $C_T = \mathrm{supp}(W_1)$ and one would have $\alpha_T = \beta_T = 0$.

The Z experiments that comprise the test are performed in a statistically independent manner. Therefore the probability, according to the hypothesis H_i, to get a sample result $\boldsymbol{X}$ equal to some element $X = (x_1, x_2, \ldots, x_Z)$ is

$$P_{H_i}(\boldsymbol{X} = X) = W_i(x_1) \cdot W_i(x_2) \cdot \cdots \cdot W_i(x_Z) \tag{3}$$

and the probability to find a result in a set $C \subset \Omega^Z$ is

$$P_{H_i}(\boldsymbol{X} \in C) = \sum_{X \in C} W_i(x_1) \cdot W_i(x_2) \cdot \cdots \cdot W_i(x_Z). \tag{4}$$

We shall call the corresponding probability assignment on Ω^Z, which is given by (4), the Zth power of W_i and write it as W_i^Z. Using the notation of measure theory one may then write the probabilities of error of first and second kind as

$$\alpha_T = W_0^Z(C_T), \tag{5}$$

$$\beta_T = W_1^Z(\Omega^Z \backslash C_T). \tag{6}$$

Now we have to define an appropriate notion of distance of probability assignments in such a way that the necessary sample size Z can be related to the distance and the error probabilities α_T and β_T. The size of the set Ω should not enter this relation. In order to simplify the problem we shall not consider Z as a function of α_T and β_T but consider an upper limit for these probabilities. It is reasonable to require that at least $\alpha_T + \beta_T < 0.5$. We shall see that the distance

$$\| W_0 - W_1 \| = \sum_{x \in \Omega} |W_0(x) - W_1(x)| \tag{7}$$

permits a judgment of this kind. Another natural distance on the space of signed measures on some σ-algebra would be

$$\| W_0 - W_1 \|_\infty = \sup_{A \in \sigma} |W_0(A) - W_1(A)|.$$

Note however, that on the subset of probability measures these notions of distance differ simply by a factor 2. One has $\| W_0 - W_1 \| = 2 \| W_0 - W_1 \|_\infty$.

3 Results

Theorem 1 *Let W_0 and W_1 be two probability assignments on a finite set Ω. There does not exist any test with sample size Z that can test the simple null hypothesis $H_0 = (W_0$ is the correct assignment) against the simple alternative $H_1 = (W_1$ is the correct assignment) and that satisfies $\alpha_T + \beta_T < 1 - Z\|W_0 - W_1\|$.*

It is useful to examine a numeric example. If we require that $\alpha_T + \beta_T < 0.5$ the theorem implies that one needs at least a sample size $Z > (2\|W_0 - W_1\|)^{-1}$ to distinguish the assignments. If $\|W_0 - W_1\| \approx 10^{-100}$, this condition is clearly prohibitive. However, 10^{-100} is still a very large number as compared to individual probabilities of micro-states of thermodynamic systems. This fact causes the instabilities of certain generalized entropies.

Proof Equations (5) and (6) imply

$$W_1^Z(C_T) - W_0^Z(C_T) = 1 - \alpha_T - \beta_T. \tag{8}$$

We have to show that there does not exist a set $C_T \subset \Omega^Z$ such that

$$W_1^Z(C_T) - W_0^Z(C_T) = 1 - \alpha_T - \beta_T > Z\|W_0 - W_1\|.$$

Or to put it differently, we have to show that

$$\forall (C \subset \Omega^Z): \quad W_1^Z(C) - W_0^Z(C) \le Z\|W_0 - W_1\|. \tag{9}$$

To that end we first prove the following lemma:

Lemma 1 *Let $v_1, v_2, \dots, v_Z$ and $\hat{v}_1, \hat{v}_2, \dots, \hat{v}_Z$ be two Z-tuples of non-negative numbers. Then*

$$\begin{aligned}\left|\prod_{i=1}^{Z} v_i - \prod_{i=1}^{Z} \hat{v}_i\right| &\le v_1 v_2 \cdots v_{Z-1}|v_Z - \hat{v}_Z| + v_1 v_2 \cdots v_{Z-2}|v_{Z-1} - \hat{v}_{Z-1}|\hat{v}_Z \\ &\quad + v_1 v_2 \cdots v_{Z-3}|v_{Z-2} - \hat{v}_{Z-2}|\hat{v}_{Z-1}\hat{v}_Z + \cdots \\ &\quad + \cdots + |v_1 - \hat{v}_1|\hat{v}_2 \hat{v}_3 \cdots \hat{v}_Z.\end{aligned}$$

Proof First we subtract and add to the difference of products $v_1 \cdots v_Z - \hat{v}_1 \cdots \hat{v}_Z$ mixed products of the kind $v_1 \cdots v_{Z-k}\hat{v}_{Z-k+1} \cdots \hat{v}_Z$:

$$\begin{aligned}v_1 \cdots v_Z - \hat{v}_1 \cdots \hat{v}_Z &= v_1 \cdots v_Z - v_1 \cdots v_{Z-1}\hat{v}_Z + v_1 \cdots v_{Z-1}\hat{v}_Z - v_1 \cdots v_{Z-2}\hat{v}_{Z-1}\hat{v}_Z \\ &\quad + v_1 \cdots v_{z-2}\hat{v}_{Z-1}\hat{v}_Z - \cdots + v_1\hat{v}_2 \cdots \hat{v}_Z - \hat{v}_1 \cdots \hat{v}_Z \\ &= v_1 \cdots v_{Z-1}(v_Z - \hat{v}_Z) + v_1 \cdots v_{z-2}(v_{Z-1} - \hat{v}_{Z-1})\hat{v}_Z + \cdots \\ &\quad + (v_1 - \hat{v}_1)\hat{v}_2 \cdots \hat{v}_Z.\end{aligned}$$

The statement of the lemma follows applying the triangle inequality. □

Corollary $\|W_0^Z - W_1^Z\| \le Z\|W_0 - W_1\|$.

Proof

$$\begin{aligned}
\|W_0^Z - W_1^Z\| &= \sum_{X\in\Omega^Z} |W_0(x_1)W_0(x_2)\cdots W_0(x_Z) - W_1(x_1)W_1(x_2)\cdots W_1(x_Z)| \\
&\le \sum_{X\in\Omega^Z} W_0(x_1)\cdots W_0(x_{Z-1})|W_0(x_Z) - W_1(x_Z)| \\
&\quad + \sum_{X\in\Omega^Z} W_0(x_1)\cdots W_0(x_{Z-2})|W_0(x_{Z-1}) - W_1(x_{Z-1})|W_1(x_Z) + \cdots \\
&\quad + \sum_{X\in\Omega^Z} |W_0(x_1) - W_1(x_1)|W_1(x_2)\cdots W_1(x_Z) \\
&= Z\|W_0 - W_1\|.
\end{aligned}$$

We note in passing that this corollary can be generalized to infinite sets Ω. □

Now all elements to complete the proof of statement (9) are ready. Thus let $C \subset \Omega^Z$ be given arbitrarily. One has

$$\begin{aligned}
W_1^Z(C) - W_0^Z(C) \le |W_1^Z(C) - W_0^Z(C)| &\le \sum_{X\in C} |W_0^Z(X) - W_1^Z(X)| \\
&\le \sum_{X\in\Omega^Z} |W_0^Z(X) - W_1^Z(X)| \le Z\|W_0 - W_1\|.
\end{aligned}$$

This completes the proof of Theorem 1. □

Theorem 2 *Let Z be a positive integer, $\varepsilon > 0$, and W_0, W_1 two probability assignments on a finite set Ω of micro-states. If $\|W_0 - W_1\| > 2(\varepsilon Z)^{-1/2}$ then there exists a set $C_T \subset \Omega^Z$ such that a test of the null hypothesis "W_0 is the correct assignment" against the alternative "W_1 is the correct assignment" that uses C_T as critical set has probabilities of error of first and second kind smaller or equal than ε: $\alpha_T \le \varepsilon$ and $\beta_T \le \varepsilon$.*

Proof Let $B = \{x \in \Omega : W_1(x) > W_0(x)\}$ and let us denote the probability, according to the hypothesis H_i, to find $x \in B$ by p_i;

$$p_i = \sum_{x\in B} W_i(x).$$

One has

$$p_1 - p_0 = \frac{1}{2}\|W_1 - W_0\|. \tag{10}$$

For $X \in \Omega^Z$ let $n(X)$ denote the number of elements of the Z-tuple $X = (x_1, \ldots, x_Z)$ that lie in B. We shall show that the following set fulfills the requirements of the critical set:

$$C_T = \left\{X \in \Omega^Z : n(X) > Zp_0 + \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}}\right\}. \tag{11}$$

We have to estimate the probabilities of the errors;

$$\alpha_T = P_{H_0}(H_0 \text{ rejected}) = P_{H_0}\left((n(\boldsymbol{X}) - Zp_0) > \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}}\right) \leq P_{H_0}\left((n(\boldsymbol{X}) - Zp_0)^2 > \frac{Z}{4\varepsilon}\right). \tag{12}$$

The variable $\xi = (n(\boldsymbol{X}) - Zp_0)^2$ has expectation value $E_0(\xi) = Zp_0(1 - p_0) \leq Z/4$. Applying Chebyshev's inequality to that stochastic variable and combining the result with inequality (12) one gets $\alpha_T \leq \varepsilon$. It remains to estimate the probability of an error of the second kind:

$$\beta_T = P_{H_1}(H_0 \text{ accepted}) = P_{H_1}\left(n(\boldsymbol{X}) \leq Zp_0 + \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}}\right).$$

According to the hypothesis of the theorem and with (10) we have

$$p_0 < p_1 - \frac{1}{\sqrt{\varepsilon Z}}$$

and hence

$$n(\boldsymbol{X}) \leq Zp_0 + \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}} \quad \Rightarrow \quad n(\boldsymbol{X}) \leq Zp_1 - Z\frac{1}{\sqrt{\varepsilon Z}} + \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}}.$$

Therefore the probability of second kind can be estimated as follows:

$$\beta_T \leq P_{H_1}\left(n(\boldsymbol{X}) \leq Zp_1 - \frac{1}{2}\sqrt{\frac{Z}{\varepsilon}}\right) \leq P_{H_1}\left((n(\boldsymbol{X}) - Zp_1)^2 \geq \frac{Z}{4\varepsilon}\right). \tag{13}$$

The expectation value of $\psi = (n(\boldsymbol{X}) - Zp_1)^2$ is $E_1(\psi) = Zp_1(1 - p_1) \leq Z/4$. Hence inequality (13) and Chebyshev's inequality imply $\beta_T \leq \varepsilon$. This concludes the proof of Theorem 2. $\square$

There is an important asymmetry between Theorems 1 and 2. The non-existence of a test to distinguish two probability assignments is independent of any physical properties of the micro-states x and obliges to consider the probability assignments as equivalent for all practical purposes. On the other hand, Theorem 2 only states the existence of a critical set, it does not assure the existence of a real test procedure. The existence of a test depends on the physical nature of the states. For instance let us consider a closed mechanical system with mixing dynamics that obeys Liouville's theorem. Let us take W_0 as the micro-canonical equilibrium distribution on phase space; $W_0(x) = 1/n$, where n is the total umber of accessible phase cells. Now consider a distribution $\tilde{W}_1$ which describes a thermodynamic state far off the equilibrium and let W_1 be the probability assignment that evolves from $\tilde{W}_1$ by means of the area conserving, mixing dynamics of the system. Obviously there does exists a critical set in Ω^Z with fairly small Z (in fact $Z = 1$ would be enough) to distinguish W_1 from W_0. But, because of the mixing property of dynamics, this set would not be accessible to real physical measurements and physically W_1 would be equivalent to the equilibrium state despite the fact that $\|W_0 - W_1\| \approx 2$.

4 Other Norms on Probability Spaces

In the previous section it was shown that the l_1-norm provides an adequate instrument to measure whether two probability assignments are too close to each other in order to be considered as different. One may now ask the question whether other l_p-norms with $p > 1$ may also be used for that purpose. The definition of such norm is given as follows:

$$\|W\|_p = \left(\sum_{x\in\Omega}(W(x))^p\right)^{1/p}. \tag{14}$$

Although all norms on finite dimensional vector spaces are topologically equivalent, in relation to the present issue they behave differently. In fact this is not surprising. As the question of whether two probability assignments are distinguishable was formulated in a way independent of the number of elements of Ω, the finite dimensional spaces in question have to be looked upon as subspaces of an infinite dimensional space of sequences, and in these spaces the norms are no longer equivalent.

To begin with, other l_p-norms are inconvenient because the norm of probability assignments varies. On the other hand, the l_1-norm of all probability assignments has the same value, which is 1. A criterion for closeness of two assignments formulated with a l_p-norm should therefore be given in a relative manner using a quantity of the kind

$$C(W_0, W_1) = \frac{\|W_0 - W_1\|_p}{\|W_0\|_p + \|W_1\|_p}. \tag{15}$$

We shall show, with a counterexample, that such a kind of closeness measure does not provide a criterion to judge whether two assignments are different. Let the space of microstates Ω be composed of three non-empty and disjoint subsets $\Omega = \Xi \cup A \cup B$ and let A and B have the same number of elements $|A| = |B| = k$. Let V be a probability assignment with support on Ξ. Then let us consider the following two probability assignments

$$W_0(x) = \begin{cases} \frac{1}{2}V(x) & \text{if } x \in \Xi, \\ \frac{1}{2k} & \text{if } x \in A, \\ 0 & \text{if } x \in B \end{cases} \quad \text{and} \quad W_1(x) = \begin{cases} \frac{1}{2}V(x) & \text{if } x \in \Xi, \\ 0 & \text{if } x \in A, \\ \frac{1}{2k} & \text{if } x \in B. \end{cases}$$

No matter how large k might be it is obviously possible to find critical sets in Ω^Z with fairly small values of Z to distinguish these two probability assignments with satisfactory reliability, in fact the l_1-distance $\|W_0 - W_1\|$ is 1 for all k. Let us now suppose that $|\Xi| \ll k$ and examine the limit of the quantity C of (15). Supposing $p > 1$ we have

$$\|W_0\|_p = \|W_1\|_p = (2^{-p}\|V\|_p^p + 2^{-p}k^{1-p})^{1/p} \underset{k\to\infty}{\to} \frac{1}{2}\|V\|_p,$$

$$\|W_0 - W_1\|_p = (2^{1-p}k^{1-p})^{1/p} \underset{k\to\infty}{\to} 0.$$

Hence the closeness measure as defined by (15) can become arbitrarily small but the two probability assignments are clearly distinguishable. Thus the l_p-norm with $p > 1$ is not appropriate to judge closeness of probability assignments.

5 Information Theoretic Closeness of Assignments

Information theoretic functionals such as the Kullback-information divergence may be considered a natural way to judge closeness of probability assignments. Therefore it is interesting to see whether analogues of the Theorems 1 and 2 may be formulated with the Kullback divergence; $I(W_2 \parallel W_1) = \sum_x W_2(x) \ln(W_2(x)/W_1(x))$. In fact the Csiszar–Kullback inequality [13]

$$\|W_2 - W_1\| \leq \sqrt{2I(W_2 \parallel W_1)}$$

guarantees that an analogue of Theorem 1 can be formulated with the Kullback-information divergence. On the other hand, as I may be arbitrarily large no matter how close the assignments are, there is no analogue to Theorem 2. Of course if one finds an experimental outcome $\boldsymbol{x} = x^*$ with $W_1(x^*) = 0$ and $W_2(x^*) \neq 0$, which could happen if $I = \infty$, one would know without any doubt that W_2 is the correct assignment. But if $\|W_2 - W_1\|$ is extremely small the probability of such a lucky event would be correspondingly small. That means $I = \infty$ does not guarantee that we may distinguish the assignments in practical terms even if we suppose that all types of sets of micro-states are physically accessible.

6 Conclusion

It was shown that the natural way to measure how close one probability assignment is to an other one is given by the l_1-distance. If this distance is smaller than $(2Z)^{-1}$ it is impossible to find a reasonably reliable test with sample size Z to distinguish the assignments. On the other hand, if the distance is larger than $2(\varepsilon Z)^{-1/2}$ there exists a critical set in the space of Z-tuples of results such that the probabilities of error of first and second kind become smaller or equal ε. Further it was shown that other l_p-norms do not permit this kind of judgment. Kullback-information divergence can also be used to give a criterion when two assignments are indistinguishable.

Acknowledgement We thank C. Tsallis for stimulating discussion.

References

1. Abe, S.: Phys. Rev. E **66**, 046134 (2002)
2. Souza, A.M.C., Tsallis, C.: Phys. Lett. A **319**, 273 (2003)
3. Lesche, B.: Phys. Rev. E **70**, 017102 (2004)
4. Naudts, J.: Rev. Math. Phys. **16**, 809 (2004)
5. Curado, E.M.F., Nobre, F.D.: Phys. A **335**, 94 (2004)
6. Kaniadakis, G., Scarfone, A.M.: Phys. A **340**, 102 (2004)
7. Abe, A., Kaniadakis, G., Scarfone, A.M.: J. Phys. A: Math. Gen. **37**, 10513 (2004)
8. Abe, S.: Phys. Rev. E **72**, 028102 (2005)
9. Ruch, E.: Theor. Chim. Acta **38**, 167 (1975)
10. Lesche, B.: J. Stat. Phys. **27**, 419 (1982)
11. Reif, F.: Fundamentals of Statistical and Thermal Physics. McGraw–Hill, New York (1965)
12. Lindgren, B.W.: Statistical Theory, 4th edn. Chapman & Hall, New York (1993)
13. Csiszár, I.: Stud. Sci. Math. Hung. **2**, 299 (1967)