On Variational Definition of Quantum Entropy

Roman V. Belavkin

School of Science and Technology, Middlesex University, London NW4 4BT, UK

Abstract. Entropy of distribution P can be defined in at least three different ways: 1) as the expectation of the Kullback-Leibler (KL) divergence of P from elementary δ-measures (in this case, it is interpreted as expected surprise); 2) as a negative KL-divergence of some reference measure ν from the probability measure P; 3) as the supremum of Shannon's mutual information taken over all channels such that P is the output probability, in which case it is dual of some transportation problem. In classical (i.e. commutative) probability, all three definitions lead to the same quantity, providing only different interpretations of entropy. In non-commutative (i.e. quantum) probability, however, these definitions are not equivalent. In particular, the third definition, where the supremum is taken over all entanglements of two quantum systems with P being the output state, leads to the quantity that can be twice the von Neumann entropy. It was proposed originally by V. Belavkin and Ohya [1] and called the *proper* quantum entropy, because it allows one to define quantum conditional entropy that is always non-negative. Here we extend these ideas to define also quantum counterpart of proper cross-entropy and cross-information. We also show inequality for the values of classical and quantum information.

Keywords: von Neumann entropy; Quantum information; Entanglement; Quantum channel
PACS: 03.65.Ca; 03.67.-a; 03.67.Mn

INTRODUCTION

Quantum probability [2] is a non-commutative generalisation of the classical probability theory [3]. Thus, the latter is a proper subset of the former, and it is reasonable to expect that any concept in quantum probability should reduce to its classical counterpart once the commutativity condition is imposed. For example, the definitions of quantum entropy and quantum information should agree with the classical definitions in the commutative case. The simplest way to achieve this is by defining quantum concepts by analogy with the classical ones and performing only minimal and necessary adjustments. For example, the definition $S[p] := -\mathrm{tr}\{(\ln p)p\}$ of the von Neumann entropy is the direct counterpart of the classical entropy $H[p] := -\sum(\ln p)p$. The information distance $D_{AU}[p,q] := \mathrm{tr}\{(\ln p - \ln q)\,p\}$ of Araki and Umegaki [4, 5] is the analogue of the classical Kullback-Leibler (KL) divergence $D_{KL}[p,q] := \sum(\ln p - \ln q)p$ [6]. One may question, however, whether such a minimalistic approach is always the right one. Non-commutativity is a subtle property having profound implications on many mathematical concepts. In this paper, we discuss an alternative definition of quantum entropy based on variational principle [1], which has a number of advantages over the von Neumann entropy. We also give new definitions of quantum cross-entropy, cross-information and prove several basic theorems.

ENTROPY IN CLASSICAL PROBABILITY

Let us review how entropy can be defined in classical probability. Consider a probability space $(\Omega, \mathscr{A}, P)$, where Ω is the set of elementary events, $\mathscr{A} \subseteq 2^{\Omega}$ is a σ-algebra of events, and $P : \mathscr{A} \to [0,1]$ is a probability measure. A random variable is a $\mathscr{A}$-measurable function $x : \Omega \to \mathbb{R}$, the *expected value* of which is the integral:

$$\mathbb{E}_P\{x\} = \int_{\Omega} x(\omega)\, dP(\omega)$$

Formally, the entropy can be defined as the expectation of $x(\omega) = -\ln p(\omega)$, where $p(\omega)$ is a P-integrable function proportional to $dP(\omega)$ (i.e. a density function):

$$H(\Omega) := H[p] = \mathbb{E}_p\{-\ln p\} = -\int_{\Omega} [\ln p(\omega)]\, dP(\omega)$$

Note that the negative logarithm of $dP(\omega)$, sometimes referred to as *surprise* associated with event ω, is the KL-divergence $D_{KL}[\delta_{\omega}, P]$ of measure P from elementary measure δ_{ω} concentrated entirely on $\omega \in \Omega$. Thus, entropy can be interpreted as a measure of expected surprise.

More generally, if P is absolutely continuous with respect to measure ν, then one can define relative entropy as negative KL-divergence $D_{KL}[P, \nu]$:

$$H[P/\nu] := -\int_{\Omega} \ln \frac{dP(\omega)}{d\nu(\omega)}\, dP(\omega) = \ln \nu(\Omega) - \int_{\Omega} \ln \frac{dP(\omega)}{dQ(\omega)}\, dP(\omega)$$

where we set $Q(E) = \nu(E)/\nu(\Omega)$ assuming $\nu(\Omega) < \infty$. When $dP/d\nu$ is proportional to dP, the relative entropy coincides with the usual definition up to an additive constant. Thus, entropy represents negative KL-divergence of some reference measure ν (e.g. Haar measure) from P.

Another way to define entropy is using information communicated between two systems. Recall that system A influences system B (or B depends on A) if the conditional probability $P(B \mid A)$ is different from the prior probability $P(B)$; or equivalently, if the joint probability $P(A \cap B)$ is different from the product probability $Q(A) \otimes P(B)$ of its marginals. This difference, measured by the KL-divergence $D_{KL}[P(A \cap B), P(A) \otimes P(B)]$, is called Shannon's *mutual information* [7]:

$$I_S(A,B) := \int_{A \times B} \left[\ln \frac{dP(b \mid a)}{dP(b)} \right] dP(a,b)$$

It is not difficult to rewrite the definition of mutual information using marginal and joint entropies

$$I_S(A,B) = H(A) + H(B) - H(A \cap B)$$

or as the difference of marginal and conditional entropies

$$I_S(A,B) = H(B) - H(B \mid A) = H(A) - H(A \mid B)$$

Mutual information is always non-negative (because the KL-divergence is) with $I_S(A,B) = 0$ if and only if A and B are independent (i.e. $P(B \mid A) = P(B)$). The supremum of $I_S(A,B)$ over all channels $P(B \mid A)$ (or $P(A \mid B)$) is attained when the channel corresponds to an injective mapping $f : A \to B$ (or $g : B \to A$), and it can be infinite. The conditional entropy $H(B \mid A)$ (or $H(A \mid B)$) in this case is zero, so that mutual information equals the marginal entropy $H(B)$ (or $H(A)$). In fact, the bounds are defined by the Shannon's inequality:

$$0 \leq I_S(A,B) \leq \min[H(A), H(B)] \tag{1}$$

For example, if $A \equiv B$, then conditional entropies are zero for any bijection $f : A \to B$, so that $I_S(A,B) = H(A) = H(B)$ is the supremum of $I_S(A,B)$. Thus, we can give the following variational definition of entropy:

$$H(B) = I_S(B,B) = \sup_{P(A \cap B)} \left\{ I_S(A,B) : \int_A dP(B \mid a)\, dQ(a) = P(B) \right\}$$

where the supremum is taken over all joint probability measures $P(A \cap B)$ such that $P(B)$ is their marginal. Observe that if one also fixes marginal $Q(A)$, then the problem is dual of the transportation problem. For example, if $A \equiv B$ and $Q(A) = P(B)$, then the solution is $P(B \mid B)$ corresponding to the identity mapping $\mathrm{id} : B \to B$. In this context, $I_S(B,B) = H(B)$ is called *self-information*. More generally, the variational definition shows that entropy $H(B)$ is an information *potential*, because it represents the maximum information that system B with distribution $P(B)$ can communicate about another system.

Despite having different interpretations the discussed above definitions of classical entropy lead to the same mathematical expression and quantity. The situation turns out to be different in quantum probability.

PROPER QUANTUM ENTROPY

Recall that in quantum (or non-commutative) probability the algebra of elementary events is defined as an algebra $\mathscr{A}(\mathscr{H})$ of subspaces $E \subseteq \mathscr{H}$ of a separable complex Hilbert space $\mathscr{H}$. Unlike algebra of subsets, this algebra is not distributive, and therefore not Boolean. It is equivalent to a non-commutative algebra of orthogonal projectors $I_E : \mathscr{H} \to \mathscr{H}$, $I_E = I_E^* = I_E^2$ onto subspaces. Instead of random variables, one considers a non-commutative $*$-algebra (an involution algebra, such as a C^* or a von Neumann algebra) of self-adjoint operators $x : \mathscr{H} \to \mathscr{H}$, $x^* = x$, which are called quantum *observables*. Instead of probability measures and their density functions, one considers operators $y : \mathscr{H} \to \mathscr{H}$, which are positive with respect to, say, trace pairing (i.e. $\langle x^* x, y \rangle = \mathrm{tr}\{x^* xy\} \geq 0$ for all x) and normalised ($\mathrm{tr}\{y\} = 1$). Such operators are called *states* or *density operators*. At this point it is important to note one crucial difference between the quantum and classical probabilities: The set $\mathscr{P}(X) := \{y : \langle x^* x, y \rangle \geq 0, \langle 1, y \rangle = 1, \ \forall x \in X\}$ of all states is *not* a simplex in quantum probability (unlike the set of all probability measures on Ω). In particular, every mixed state $p \in \mathscr{P}(X)$ can be

represented as a convex combination of extreme points $\delta \in \operatorname{ext}\mathscr{P}(X)$ in a non-unique way. This is related to the following fact.

Any Boolean subalgebra $\mathscr{C}(\mathscr{H}) \subset \mathscr{A}(\mathscr{H})$ can be identified with a commutative subalgebra of orthogonal projectors, which can be diagonalised in the same basis $\{e_i\}_{i\in\mathbb{N}} \subset \mathscr{H}$. Thus, fixing the set Ω of elementary events in classical probability is equivalent to fixing a basis $\{e_i\}_{i\in\mathbb{N}}$ in the Hilbert space $\mathscr{H}$ and considering only diagonal with respect to it operators. Thus, from a mathematical point of view, a transition from classical to quantum formalism can be seen as a relaxation of constraints (i.e. a restriction to a specific orthogonal basis). Physical motivation of this relaxation, however, is the fact that quantum objects have properties (e.g. position and momentum) that cannot be established simultaneously in any experiment (due to the uncertainty principle). Thus, quantum systems are fundamentally more 'uncertain' than classical, and therefore the definition of quantum entropy should reflect this additional and irreducible uncertainty. This can be achieved if quantum entropy is defined as the supremum of quantum mutual information, because it is taken over a larger (due to non-commutativity) set of quantum states, and this is how the *proper* quantum entropy was defined [1].

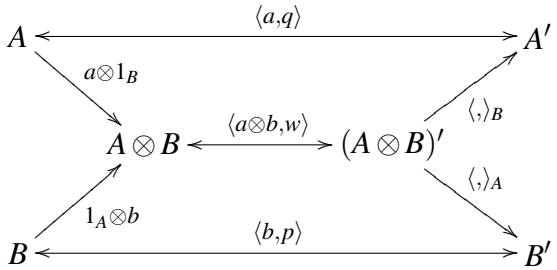
Specifically, let $A \otimes B$ be the tensor product of two algebras corresponding to two subsystems of a composite system, and let $\mathscr{P}(A \otimes B)$ be the set of all *compound states* w, which play the role of joint probability measures. Taking partial traces $q = \langle 1, w\rangle_B$ and $p = \langle 1, w\rangle_A$ one obtains states $q \in \mathscr{P}(A)$ and $p \in \mathscr{P}(B)$, called the marginal or *reduced* states of w. Compound states of the form $q \otimes p$ are called *product* states. Convex closure $\operatorname{cl}\operatorname{co}[\mathscr{P}(A) \otimes \mathscr{P}(B)]$ of all product states is the set of *separable* states. We remind that in quantum probability, there are compound states that are not separable:

$$\mathscr{P}(A \otimes B) \setminus \operatorname{cl}\operatorname{co}[\mathscr{P}(A) \otimes \mathscr{P}(B)] \neq \varnothing$$

Non-separable compound states correspond to non-classical coupling (dependency or communication) between sub-systems A and B, which is called a quantum *entanglement*. In operational theory of entanglement [1] a *generalized entanglement* of reduced states q and p associated with the compound state w is defined by normal completely positive operations $\pi : A \to \mathscr{P}(B)$ or $\pi' : B \to \mathscr{P}(A)$ defined respectively as follows:

$$\pi(a) = \langle a \otimes 1_B, w\rangle_A\,, \qquad \pi'(b) = \langle 1_A \otimes b, w\rangle_B$$

These entanglement operations are composed of several linear maps, shown on the diagram below:



It is easy to check that $\pi(1_A) = p \in \mathscr{P}(B)$, $\pi'(1_B) = q \in \mathscr{P}(A)$ and $\langle a, \pi'(b)\rangle_A = \langle a \otimes b, w\rangle = \langle b, \pi(a)\rangle_B$. It was proven in [1] that if w is a separable state, then the composition

of entanglement with transposition $a \mapsto [\pi(a)]'$ (or $b \mapsto [\pi'(b)]'$) is also completely positive. Dually, if these maps are not completely positive, then the compound state w is not separable, and the coupling is called a *proper* (or true quantum) entanglement.

Every entanglement $\pi : A \to \mathscr{P}(B)$ has decomposition (see [1]):

$$\pi(a) = p^{1/2}\Pi(a)p^{1/2}$$

where $\Pi : A \to B'$ is a normal completely positive contraction such that $1_B \geq \Pi(1_A) \geq P_p$, where $P_p \in B'$ is the minimal orthoprojector on the support of state $p \in \mathscr{P}(B)$. The entanglement of the form $\pi(a) = p^{1/2}ap^{1/2}$ for $A \subseteq B'$ is called *standard* (i.e. $\Pi(a) = a$ is an injection into B').

The compound state $w \in \mathscr{P}(A \otimes B)$ defines a channel $T : \mathscr{P}(A) \to \mathscr{P}(B)$ (a Markov morphism) transforming the reduced states $q \mapsto Tq = p$. The adjoint of T is a unital completely positive map $T^* : B \to A$. As in classical information theory, the divergence of $q \otimes p$ from w, defining the quantum channel capacity, is called the quantum mutual information:

$$I_S(A,B) := D_{AU}[w, q \otimes p] = \langle \ln w - \ln q \otimes p, w \rangle$$

It was first considered in [8]. The mutual information can be written using entropies:

$$I_S(A,B) = S(A) + S(B) - S(A \otimes B)$$

Here, $S(A) := S[q] = -\langle \ln q, q \rangle$ denotes the von Neumann entropy of state $q \in \mathscr{P}(A)$. Stretching the analogy further, one may write

$$I_S(A,B) = S(A) - S(A \mid B) = S(B) - S(B \mid A)$$

where $S(B \mid A) = S(A \otimes B) - S(A) = S(B) - I_S(A,B)$ can be seen as the quantum analogue of conditional entropy. However, such definitions lead to undesired results.

Indeed, if $w \in \mathscr{P}(A \otimes B)$ is a non-separable compound state, then the joint von Neumann entropy $S(A \otimes B)$ can be less than the marginal entropies $S(A)$ or $S(B)$. For example, w can be a pure state with non-pure marginal states p and q. In this case, $S(A \otimes B) = 0$, but $S(A) > 0$ and $S(B) > 0$. Thus, the Shannon's inequality (1) does not hold for the von Neumann entropies. In fact, quantum mutual information can be twice the von Neumann entropy (e.g. $S(A \otimes B) = 0$ and $S(A) = S(B)$, so that $I_S(A \otimes B) = 2S(A)$). Furthermore, the conditional entropy $S(B \mid A) = S(B) - I_S(A,B)$ can be negative (e.g. $S(A \otimes B) = 0$ and $S(B \mid A) = -S(A)$).

In order to reconcile quantum information theory with the classical ideas, one can use variational definition of quantum entropy of state $p \in \mathscr{P}(B)$ as the supremum of mutual information taken over all compound states $w \in \mathscr{P}(A \otimes B)$ with the marginal state $p = \langle 1, w \rangle_A$:

$$H[B] := I_S(B,B) = \sup_{w \in \mathscr{P}(A \otimes B)} \{I[w, q \otimes p] : \langle 1, w \rangle_A = p\}$$

This is equivalent to taking the supremum over entanglement operations $\pi : A \to \mathscr{P}(B)$ with the output state $p \in \mathscr{P}(B)$. The supremum is attained at $\overline{w}$ corresponding to the

standard entanglement $\pi(a) = p^{1/2}ap^{1/2}$ [1]. This definition of quantum entropy was first introduced in [1], and it was called the *proper* quantum entropy. It is greater than the von Neumann entropy and satisfies the Shannon's inequality (1). The proper quantum conditional entropy is defined by the difference $H(B \mid A) := H(B) - I_S(A,B)$, which is non-negative.

We note also that non-commutativity allows for different definitions of information distance between states. Indeed, given two states y and z, their Radon-Nikodym derivative y/z is not uniquely defined, and $\ln(y/z) \neq \ln y - \ln z$, unless y, z commute. The naive definition $y/z := \exp(\ln y - \ln z)$ corresponds to information distance $I[y,z] = \langle \ln y - \ln z, y\rangle$, which is the Araki-Umegaki information [4, 5]. Alternatively, one can use Hermitian operators $y/z := y^{1/2}z^{-1}y^{1/2}$ or $y/z := z^{-1/2}yz^{-1/2}$, which lead to different forms of additive quantum information [9]. Such a definition gives a better contrast contrast between states that do not commute [10].

QUANTUM CROSS-ENTROPY AND CROSS-INFORMATION

Other information-theoretic quantities can be defined using the von Neumann and proper quantum entropies. Thus, quantum *cross-entropy* of the von-Neumann type can be defined by analogy with the classical theory:

$$S[p,q] := -\langle \ln q, p\rangle = S[p] + I[p,q]$$

The *proper* quantum cross-entropy is defined using the proper quantum entropy as $H[p,q] := H[p] + I[p,q]$. Clearly, $H[p,q] \geq S[p,q]$.

If $A \subseteq B$ (in the sense that there is an injection $f : A \to B$), then state $q \in \mathscr{P}(A)$ on A can also be considered as a state on B (i.e. as $p = q \circ f^{-1} \in \mathscr{P}(B)$). Thus, we can consider product state $q \otimes q \in \mathscr{P}(A \otimes B)$. The *cross-information* of a quantum channel $T : \mathscr{P}(A) \to \mathscr{P}(B)$ associated with compound state $w \in \mathscr{P}(A \otimes B)$ and reduced state $q = \langle 1, w\rangle_B$ is the following quantity:

$$I[w, q \otimes q] = \langle \ln w - \ln q \otimes q, w\rangle \tag{2}$$

It was introduced in [11, 12] after the observation that the triangle $(w, q \otimes q, q \otimes p)$ is always right.

Theorem 1 (Shannon-Pythagorean theorem [12])**.** *Let $w \in \mathscr{P}(A \otimes B)$, $A \subseteq B$, and let $q = \langle 1, w\rangle_B$, $p = \langle 1, w\rangle_A$. Then*

$$I[w, q \otimes q] = I[w, q \otimes p] + I[p, q]$$

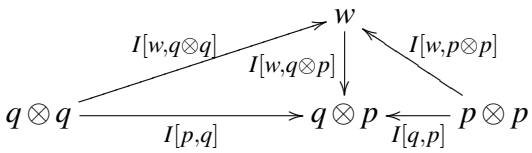
Proof. Consider the law of cosines for w, $q \otimes p$ and $q \otimes q$:

$$I[w, q \otimes q] = I[w, q \otimes p] + I[q \otimes p, q \otimes q] - \langle \ln q \otimes p - \ln q \otimes q, q \otimes p - w\rangle$$

The latter member is always zero. Indeed, $\ln q \otimes p - \ln q \otimes q = 1_A \otimes (\ln p - \ln q)$, and because $\langle 1, w\rangle_A = \langle 1, q \otimes p\rangle_A = p$, this gives $\langle 1_A \otimes (\ln p - \ln q), q \otimes p - w\rangle = 0$.

The second member $I[q\otimes p, q\otimes q] = \langle 1_A \otimes (\ln p - \ln q), q\otimes p\rangle$, which equals $I[p,q] = \langle \ln p - \ln q, p\rangle$ (for $\langle 1_A, q\rangle = 1$). □

The second cross-information $I[w, p\otimes p]$ associated with $w \in \mathscr{P}(A\otimes B)$ and $p = \langle 1, w\rangle_A$ is defined similarly: $I[w,p\otimes p] = I[w,q\otimes p] + I[q,p]$. Geometric interpretation of cross-information as the hypotenuse of the right triangle $(w, q\otimes q, q\otimes p)$ is shown on the diagram below:



The arrows represent the idea that compound state $w \in \mathscr{P}(A\otimes B)$ defines channels $T : \mathscr{P}(A) \to \mathscr{P}(B)$ or $T^{-1} : \mathscr{P}(B) \to \mathscr{P}(A)$ transforming $q \mapsto Tq = p$ or $p \mapsto T^{-1}p = q$.

Corollary. *For $w \in \mathscr{P}(A\otimes B)$, $q = \langle 1, w\rangle_B$, $p = \langle 1, w\rangle_A$:*

$$I[p,q] \le I[q,p] \quad \Longleftrightarrow \quad I[w, q\otimes q] \le I[w, p\otimes p]$$

Proof. Follows from the fact that mutual information equals to the following two differences: $I[w,q\otimes p] = I[w,q\otimes q] - I[p,q] = I[w,p\otimes p] - I[q,p]$. □

Corollary. *Cross-information $I[w,q\otimes q]$ is the difference of cross-entropy and conditional entropy:*

$$I[w,q\otimes q] = S[p,q] - S(B\mid A) = H[p,q] - H(B\mid A)$$

Proof. Substitute $I[p,q] = \langle \ln p, p\rangle - \langle \ln q, p\rangle$ into $I[w,q\otimes q] = I[w,q\otimes p] + I[p,q]$:

$$I[w,q\otimes q] = \underbrace{-\langle \ln q, p\rangle}_{S[p,q]} - \underbrace{[-\langle \ln p, p\rangle - I[w,q\otimes p]]}_{S(B|A)}$$

The difference $S[p,q] - S(B\mid A)$ of the von-Neumann type entropies is equal to the difference $H[p,q] - H(B\mid A)$ of proper quantum entropies. □

One can see from Corollary that cross-information is bounded above by the proper quantum cross-entropy: $I[w,q\otimes q] \le H[p,q]$. The following inequality gives a tighter bound.

Theorem 2 (Cross-information inequality)**.**

$$I[w,q\otimes q] \le \min\{H[q], H[p]\} + I[p,q] \le H[p,q]$$

Proof. The first inequality follows from $I[w,q\otimes q] = I[w,q\otimes p] + I[p,q]$ (Theorem 1) and Shannon's inequality $I[w,q\otimes p] \leq \min\{H[q],H[p]\}$ for proper quantum entropies. The second inequality follows from $H[p] + I[p,q] = H[p,q]$. □

If $q \in \mathscr{P}(A)$ is an initial state with known entropy, then optimisation of transformations $q \mapsto Tq = p$ with respect to some utility operator may correspond to either an increase or decrease of cross-entropy $H[p,q]$ relative to $H[q]$. In particular, $H[p,q] \leq H[q]$ implies $I[w,q\otimes q] \leq H[q]$ by Theorem 2, and the following relation can be useful.

Theorem 3 (Entropy bound)**.**

$$I[w,q\otimes q] \leq H(A) \quad \Longleftrightarrow \quad I[p,q] \leq H(A \mid B)$$

Proof. Subtracting $I[w,q\otimes p]$ from both sides of inequality $I[w,q\otimes q] \leq H[q] =: H(A)$, one obtains $I[w,q\otimes q] - I[w,q\otimes p] = I[p,q]$ on the left, and conditional entropy $H[q] - I[w,q\otimes p] = H(A \mid B)$ on the right. □

REFERENCES

1. V. P. Belavkin, and M. Ohya, *Royal Society of London Proceedings Series A* **458** (2002).
2. J. von Neumann, *Mathematische Grundlagen der Quantenmechanik. (German) [Mathematical Foundations of Quantum Mechanics]*, Springer-Verlag, Berlin, 1932.
3. A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Julius Springer, Berlin, 1933, in German.
4. H. Araki, *Publications of the Research Institute for Mathematical Sciences* **11**, 809–833 (1975).
5. H. Umegaki, *Kodai Mathematical Seminar Reports* **14**, 59–85 (1962).
6. S. Kullback, and R. A. Leibler, *The Annals of Mathematical Statistics* **22**, 79–86 (1951).
7. C. E. Shannon, *Bell System Technical Journal* **27**, 379–423 and 623–656 (1948).
8. R. L. Stratonovich, *Izvestia Vuzov: Radiophysics* **4**, 15–24 (1965), in Russian.
9. V. P. Belavkin, and P. Staszewski, *Reports in Mathematical Physics* **20**, 373–384 (1984).
10. F. Hiai, and D. Petz, *Communications in Mathematical Physics* **143**, 99–114 (1991).
11. R. V. Belavkin, "Minimum of information distance criterion for optimal control of mutation rate in evolutionary systems," in *Quantum Bio-Informatics V*, edited by L. Accardi, W. Freudenberg, and M. Ohya, World Scientific, 2013, vol. 30 of *QP-PQ: Quantum Probability and White Noise Analysis*, pp. 95–115.
12. R. V. Belavkin, "Law of Cosines and Shannon-Pythagorean Theorem for Quantum Information," in *Geometric Science of Information*, edited by F. Nielsen, and F. Barbaresco, Springer, Heidelberg, 2013, vol. 8085 of *Lecture Notes in Computer Science*, pp. 369–376.