Law of Cosines and Shannon-Pythagorean Theorem for Quantum Information[*]

Roman V. Belavkin[1]

School of Engineering and Information Sciences
Middlesex University, London NW4 4BT, UK

Abstract. The concept of information distance in non-commutative setting is re-considered. Additive information, such as Kullback-Leibler divergence, is defined using convex functional with gradient having the property of homomorphism between multiplicative and additive subgroups. We review several geometric properties, such as the logarithmic law of cosines, Pythagorean theorem and a lower bound given by squared Euclidean distance. We also prove a special case of Pythagorean theorem for Shannon information, which finds applications in information-theoretic variational problems.

1 Introduction

Kullback-Leibler (KL) divergence [1] is, perhaps, the most important functional in information theory and statistics. If p and q are two probability measures defined on the same σ-algebra, and p is absolutely continuous with respect to q, then KL-divergence of q from p is defined as the expected value of $\ln(p/q)$ with respect to p:

$$D_{KL}(p,q) := \mathbb{E}_p\{\ln(p/q)\} \tag{1}$$

This functional plays a role of a distance, because $D_{KL}(p,q) \geq 0$ for all p, q, and $D_{KL}(p,q) = 0$ if and only if $p = q$, but it is not a metric (e.g. $D_{KL}(p,q) \neq D_{KL}(q,p)$ in general). In fact, it is well-known that KL-divergence plays a role of half of squared Euclidean distance [2], and it is related to the unique Riemannian metric on the set of all probability measures on the same σ-algebra [3]. In physics, KL-divergence is used to define relative entropy, and similar functionals have been used in non-commutative (or quantum) probability [4–6].

In the next section, we overview the common algebraic structure and duality used both in classical and quantum probability. We then use ideas of convex analysis to define information distance by restricting its gradient to a logarithmic map that has the property of homomorphism from multiplicative to additive subgroups of the dual spaces. Alternatively, one can restrict the gradient of the dual functional to an exponential map. In classical probability, these definitions are equivalent leading to the familiar KL-divergence. In the non-commutative case, different definitions are known to lead to different types of quantum information. Our definition corresponds to the Araki-Umegaki type [5, 4], and it

has the advantage that various geometric properties, such as law of cosines and Pythagorean theorem [2], hold for all, not necessarily normalized measures or operators. The relation to squared Euclidean distance is shown by an inequality. We give new and simple proofs of these properties. In the final section, we show that a certain orthogonality condition always holds for a triangle, defined by a joint state and its reduced (i.e. marginal) states. The corresponding Pythagorean theorem gives a new, geometric interpretation of Shannon information, and can be used to relate different information-theoretic variational problems.

2   Algebra of Random Variables and Module of Measures

Throughout this paper we shall consider two linear spaces X and Y put in duality via a non-degenerate bilinear form $\langle \cdot, \cdot \rangle : X \times Y \to \mathbb{C}$. Space X is in fact a $*$-algebra with associative, but generally non-commutative multiplication $\cdot : X \times X \to X$ and involution as a self-inverse, antilinear map $* : X \to X$ reversing the multiplication order: $(x^*z)^* = z^*x$. It is usually assumed that the algebra has a multiplicative unit $1 \in X$. Commutative algebra X is understood as an algebra of classical random variables (i.e. measurable functions with pointwise multiplication). In this case, the pairing $\langle x, y \rangle$ is understood either as the sum $\sum x_i y_i$ or the integral $\int x \, dy$. Non-commutative algebra X is understood as an algebra of Hermitian operators on a separable Hilbert space, often referred to as *observables* in physics. In this case, $\langle x, y \rangle$ is understood as trace pairing $\operatorname{tr} \{xy\}$.

The dual space Y is closed under the transposed involution $* : Y \to Y$, defined by $\langle x, y^* \rangle = \langle x^*, y \rangle^*$. If the pairing $\langle \cdot, \cdot \rangle$ has the property that for each $z \in X$ there exists a transposed element $z' \in Y$ such that $\langle zx, y \rangle = \langle x, z'y \rangle$, then Y is a left (resp. right) *module* over algebra $X \subseteq Y$ with respect to the transposed left (resp. right) action $y \mapsto z'y$ (resp. $y \mapsto yz^{*\prime*}$) of X on Y such that $(xz)' = z'x'$ and $\langle x, yz^{*\prime*} \rangle = \langle x^*, z^{*\prime}y^* \rangle^* = \langle z^*x^*, y^* \rangle^* = \langle xz, y \rangle$.

The sets of positive elements in X and Y are defined respectively as

$$X_+ := \{x : z^*z = x, \ \exists z \in X\}, \qquad Y_+ := \{y : \langle x, y \rangle \geq 0, \ \forall x \in X_+\}$$

Elements of the cone Y_+ are understood as positive measures or their non-commutative generalizations. The base of the positive cone

$$\mathcal{P}(X) := \{p \in Y_+ : \langle 1, p \rangle = 1\}$$

is a (weakly) compact convex set. When X is a commutative algebra, $\mathcal{P}(X)$ is a simplex — each $p \in \mathcal{P}(X)$ is uniquely represented as a convex combination of the extreme points δ of $\mathcal{P}(X)$. In this case, elements of $\mathcal{P}(X)$ are classical probability measures. Their non-commutative generalizations are also referred to as quantum *states*. Note that although $\langle x, y \rangle \in \mathbb{C}$ in general, the expected values $\mathbb{E}_p\{x\} = \langle x, p \rangle$ are always real, because x is self-adjoint and p is positive (and hence also self-adjoint).

One can see that random variables (observables) $x \in X$ and measures (states) $y \in Y$ can be equipped with a rich algebraic structure. This allows us, in partic-

ular, to define the exponential and logarithmic mappings by the power series:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \qquad \ln y = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} (y-1)^n$$

The exponential and logarithmic mappings are homomorphisms of additive and multiplicative groups of commutative subalgebras of X: $e^{u+v} = e^u e^v$, $\ln(yz) = \ln y + \ln z$ iff $uv = vu$ and $yz = zy$. Furthermore, if X is a tensor product $A \otimes B$ of two algebras, then $e^{u \oplus v} = e^u \otimes u^v$ and $\ln(y \otimes z) = \ln y \oplus \ln z$, where $u \oplus v := u \otimes 1_B + 1_A \otimes v$. Note that duality of X and Y allows us to consider the exponent as a mapping from X into its (pre)-dual Y, and the logarithm as its inverse. In the next section we shall consider them as gradients of dual convex functionals. Their algebraic properties are key to additivity of information.

3 Information Distance for Measures

Additivity of information from independent sources is considered to be one of the most important properties of information, and is often required as an axiom [7]. Because joint probability distributions of independent random variables are products of the marginal distributions, this leaves us little choice but to use the logarithmic mapping, which has the required property of homomorphism from multiplicative to an additive group. More specifically, we shall define information using a functional on the space of measures Y, such that its gradient at y is the logarithm of y. Alternatively, because gradients of dual convex functionals are inverse mappings of each other, we can define information as the Legendre-Fenchel dual of a functional on the space of random variables X, the gradient of which at x is the exponent of x. Let us now make this explicit.

Consider the following functional $F(y) := \langle \ln y - 1, y \rangle$. We define $(\ln 0) \cdot 0 = 0$ and $F(y) = \infty$ if $y < 0$, so that F is a proper closed convex functional. It is not difficult to check that its gradient $\nabla F : Y \to X$ is $\nabla F(y) = \ln y$, the logarithm of y. The inverse mapping $y = e^x$ is the gradient $\nabla F^* : X \to Y$ of the dual functional $F^*(x) := \langle 1, e^x \rangle$, which is the Legendre-Fenchel transform of F (i.e. $F^*(y) = \sup[\langle x, y \rangle - F(x)]$). We note also that $F(y)$ is in turn the dual of $F^*(x)$, because $F^{**} = F$ for proper closed convex functions [8].

An *information distance* $I : Y \times Y \to \mathbb{R} \cup \{\infty\}$ between two measures y and $z \in Y$ can be defined in a similar way [9]:

$$I(y, z) = \langle \ln y - \ln z, y \rangle - \langle 1, y - z \rangle \tag{2}$$

In addition to the rule $(\ln 0) \cdot 0 = 0$, we define $I(y, z) = \infty$ if $y < 0$ or $z \leq 0$. One can see that $I(p, q) = D_{KL}(p, q)$ for probability measures, as $\langle 1, p - q \rangle = 0$. Unlike the KL-divergence, however, functional (2) is non-negative for all, not necessarily normalized or positive measures. Indeed, $I(y, z) \geq 0$ for all $y, z \in Y$ and $I(y, z) = 0$ iff $y = z$. The information distance is not symmetric, as $I(y, z) \neq I(z, y)$ in general. The symmetrized distance can be defined

$$SI(y, z) = I(y, z) + I(z, y) = \langle \ln y - \ln z, y - z \rangle \tag{3}$$

Remark 1. In classical probability, the KL-divergence can be equivalently defined using $\ln(y/z)$, where y/z is the Radon-Nikodym derivative. In quantum probability, however, $\ln(y/z) \neq \ln y - \ln z$ if density operators y, z do not commute. Furthermore, if $yz \neq zy$, then the Radon-Nikodym derivative y/z is not uniquely defined. The naive definition $y/z := \exp(\ln y - \ln z)$ corresponds to quantum relative information of Araki-Umegaki [5, 4], similar to equation (2). Alternatively, one can use Hermitian operators $y/z := y^{1/2}z^{-1}y^{1/2}$ or $y/z := z^{-1/2}yz^{-1/2}$, and they lead to different forms of quantum relative information [6] known to give better contrast for non-commutative measures [10]. Finally, the definition of information as the Legendre-Fenchel dual of functional $I^*(x, z)$, the gradient of which is $e^{x/2}ze^{x/2}$ or $z^{1/2}e^x z^{1/2}$, leads to other (thermodynamic) types of quantum relative information.

Functional $I(y, z)$ is proper, closed and strictly convex, and it is the dual of functional $I^*(x, z) := \langle 1, e^x z \rangle$. Furthermore, $I(y, z)$ is Gâteaux differentiable for all y where it is finite, and its gradient has a convenient form $\nabla_y I(y, z) = \ln y - \ln z$. The second derivative $\nabla_y^2 I(y, z)$ (the Hessian) is the inverse y^{-1}, and it is defined for y with a strictly positive spectrum.

Matrix representation of operator $\nabla_p^2 I(p, q) = p^{-1}$ taken at $p = q$ is known as *Fisher information* matrix at q (note that Fisher information is usually defined in terms of partial derivatives over some vector of parameters under suitable differentiability assumptions). It plays an important role in information geometry defining the unique Riemannian metric on the set $\mathcal{P}(X)$ of all probability measures [3]. In fact, the following theorems show that information distance $I(y, z)$ has many properties similar to squared Euclidean distance.

4 Logarithmic Law of Cosines and Pythagorean Theorem

The non-symmetric version of the Pythagorean theorem for the KL-divergence of classical probability measures was first proven in [2], in which the analogue of the law of cosines was also obtained. We now present another, perhaps, more simple proof of these facts using Taylor expansion of information distance (2), the definition of which also allows us to extend these theorems to all measures or operators, not necessarily with equal norms.

Theorem 1 (Logarithmic law of cosines). *Finite information distances between any three points w, y, z in Y satisfy the relation*

$$I(w, z) = I(w, y) + I(y, z) - \langle \ln y - \ln z, y - w \rangle$$

Proof. Consider first order Taylor expansion of $I(\cdot, z)$ at y:

$$I(w, z) = I(y, z) + \langle \ln y - \ln z, w - y \rangle + R_1(y, w)$$

The law follows from the fact that the remainder $R_1(y, w) = I(w, y)$. Indeed,

$$
\begin{aligned}
R_1(y, w) &= I(w, z) - I(y, z) - \langle \ln y - \ln z, w - y \rangle \\
&= \langle \ln w - \ln z, w \rangle - \langle 1, w - z \rangle - \\
&\quad \langle \ln y - \ln z, y \rangle + \langle 1, y - z \rangle - \langle \ln y - \ln z, w - y \rangle \\
&= \langle \ln w - \ln y, w \rangle - \langle 1, w - y \rangle
\end{aligned}
$$

$\square$

Remark 2. If the information distance $I(w, y)$ is defined using the logarithm of the Radon-Nikodym derivative w/y, then proof of equality $R_1(y, w) = I(w, y)$ requires the condition $\ln(w/z) + \ln(z/y) = \ln(wz^{-1}zy^{-1}) = \ln(w/y)$, which holds if and only if $wy = yw$. Definition (2) using $\ln w - \ln y$ avoids this complication.

Remark 3 (Log-orthogonality). Let us define the squared 'Euclidean' distance:

$$
s^2(y, z) := \langle 1, (y - z)^2 \rangle \tag{4}
$$

Here we use the notation $\langle 1, y^2 \rangle := \langle y', y \rangle$. The gradient of s^2 is $\nabla_y s^2(y, z) = 2(y - z)$, and the classical law of cosines can be obtained as in Theorem 1 using the first order Taylor expansion of $s^2(w, z)$ at y:

$$
s^2(w, z) = s^2(w, y) + s^2(y, z) - 2\langle y - z, y - w \rangle
$$

When vectors $y - z$ and $y - w$ are orthogonal, then $\langle y - z, y - w \rangle = 0$, and we obtain the familiar Pythagorean theorem: $s^2(w, z) = s^2(w, y) + s^2(y, z)$. In information geometry, orthogonality of $y - z$ and $y - w$ is defined by the condition $\langle \ln y - \ln z, y - w \rangle = 0$. Observe that this condition is not symmetric. The symmetrized log-orthogonality follows from the logarithmic law of cosines for symmetrized information distance:

$$
SI(w, z) = SI(w, y) + SI(y, z) - [\langle \ln y - \ln z, y - w \rangle + \langle \ln y - \ln w, y - z \rangle]
$$

For completeness, let us include analogue of the Pythagorean theorem [2].

Corollary 1 (Log-Pythagorean theorem). *Given three points w, y, z in Y satisfying the log-orthogonality condition $\langle \ln y - \ln z, y - w \rangle = 0$, their information distances satisfy the relation: $I(w, z) = I(w, y) + I(y, z)$.*

Proof. Follows immediately from Theorem 1. $\square$

5 Inequality for Information and Euclidean Distances

The fact that KL-divergence $D(p, q)$ plays the role of half of squared Euclidean distance $s^2(p, q)$ between probability measures is well-known [2]. Another fact, perhaps less known, is the inequality $D_{KL}(p, q) + D_{KL}(q, p) \geq s^2(p, q)$ (Theorem 1.14 in [11]). We now generalize this result for information distance (2), and also give a different proof.

Theorem 2. *Information distance between y, z in Y satisfies the inequality*

$$I(y, z) \geq \frac{s^2(y, z)}{2 \max\{\|y\|_\infty, \|z\|_\infty\}}$$

where $\|\cdot\|_\infty$ is the supremum norm, and $s^2(y, z) = \langle 1, (y - z)^2 \rangle$.

Proof. Consider the first order Taylor expansion of $I(y, w)$ at z:

$$I(y, w) = I(z, w) + \langle \nabla_y I(z, w), y - z \rangle + R_1(z, y)$$

where the remainder is

$$
\begin{aligned}
R_1(z, y) &= \int_0^1 (1 - t) \left\langle 1, \nabla_y^2 I(z + t(y - z), w)(y - z)^2 \right\rangle dt \\
&= \frac{1}{2} \left\langle 1, \nabla_y^2 I(\xi, w)(y - z)^2 \right\rangle \qquad \text{for some } \xi \in [z, y) \\
&= \frac{1}{2} \left\langle \xi^{-1}, (y - z)^2 \right\rangle
\end{aligned}
$$

Here we used the notation $\langle 1, Ay^2 \rangle := \langle Ay, y \rangle$. In the second line, we used the fact that I is sufficiently smooth to apply the mean-value theorem and wrote the remainder in the Lagrange form for some ξ in the open interval $[z, y)$. Finally, we used the fact that $\nabla_y^2 I(y, w) = y^{-1}$. The result follows from the fact that $R_1(z, y) = I(y, z)$, obtained earlier in Theorem 1, and from inequality $\|\xi\|_\infty \leq \max\{\|y\|_\infty, \|z\|_\infty\}$ for any $\xi \in [z, y) = \{z + t(y - z) : t \in [0, 1)\}$. $\square$

Corollary 2. $I(p, q) + I(q, p) \geq s^2(p, q)$ *for all $p, q \in \mathcal{P}(X)$.*

Proof. Recall that $\mathcal{P}(X) := \{p : p \geq 0, \langle 1, p \rangle = 1\}$, so that $\|p\|_\infty \leq 1$ for any $p, q \in \mathcal{P}(X)$. Thus, $I(p, q) \geq \frac{1}{2}\langle 1, (p - q)^2 \rangle = \frac{1}{2}s^2(p, q)$ by Theorem 2, and the same is true for $I(q, p)$. The result is the sum of two inequalities. $\square$

6   Shannon-Pythagorean Theorem

Let algebra X be a tensor product $A \otimes B$ of two algebras corresponding to two subsystems of a composite system. We denote by 1_A and 1_B the units of these algebras. As before, Y will denote the (pre)-dual space of $X = A \otimes B$, and $\mathcal{P}(A \otimes B)$ is the set of all positive linear functionals $p(x) = \langle x, p \rangle$ preserving the unit $\langle 1, p \rangle = 1$, and they correspond to joint probability measures or joint states. If $q \in \mathcal{P}(A)$ is a state on A and $p \in \mathcal{P}(B)$ is a state on B, then $q \otimes p$ is a *product* state on $A \otimes B$. Convex closure of all product states is the set of all *separable* states. We remind that in quantum probability, there are joint states $w \in \mathcal{P}(A \otimes B)$ that are not separable. Given a joint state w on $A \otimes B$, the partial traces $q = \langle 1, w \rangle_B$ and $p = \langle 1, w \rangle_A$ are states on A and B respectively.

Lemma 1. *Let $w \in \mathcal{P}(A \otimes B)$, $z \in \mathcal{P}(A)$, $p, q \in \mathcal{P}(B)$, and let $p = \langle 1, w \rangle_A$. Then vectors $z \otimes p - z \otimes q$ and $z \otimes p - w$ are orthogonal in the following sense:*

$$\langle \ln z \otimes p - \ln z \otimes q, z \otimes p - w \rangle = 0$$

Proof. Using the property $\ln z \otimes y = \ln z \oplus \ln y = \ln z \otimes 1_B + 1_A \otimes \ln y$ we obtain

$$\ln z \otimes p - \ln z \otimes q = 1_A \otimes (\ln p - \ln q)$$

The result follows from the condition that the expectation $\langle b, p \rangle$ of observable $b \in B$ is invariant under the embedding $B \ni b \mapsto 1_A \otimes b \in A \otimes B$, if $p \in \mathcal{P}(B)$ is the reduced state of joint state $w \in \mathcal{P}(A \otimes B)$. Therefore, $\langle 1_A \otimes b, w \rangle = \langle 1_A \otimes b, z \otimes p \rangle = \langle b, p \rangle$, because $\langle 1, w \rangle_A = \langle 1, z \otimes p \rangle_A = p$. $\qquad\square$

A joint state $w \in \mathcal{P}(A \otimes B)$ defines a channel $T : \mathcal{P}(A) \to \mathcal{P}(B)$ (Markov morphism or operation) transforming $q \in \mathcal{P}(A)$ into $p = Tq \in \mathcal{P}(B)$, where $q = \langle 1, w \rangle_B$, $p = \langle 1, w \rangle_A$ are reduced states of w (i.e. marginals). The information distance $I(w, q \otimes p)$ is Shannon information defining the capacity of this channel. If $A \subseteq B$, then we may consider q also as an element of $\mathcal{P}(B)$ and consider information distance $I(p, q) = I(Tq, q)$. It turns out that $I(w, q \otimes p)$ and $I(p, q)$ are related by the Pythagorean theorem.

Theorem 3 (Shannon-Pythagorean theorem). *Let w be a joint state on $A \otimes B$, $A \subseteq B$, and let $q = \langle 1, w \rangle_B$, $p = \langle 1, w \rangle_A$ be its marginal states. Then*

$$I(w, q \otimes q) = I(w, q \otimes p) + I(p, q)$$

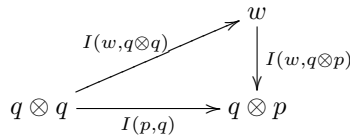*Proof.* According the law of cosines for w, $q \otimes p$ and $q \otimes q$ we have:

$$I(w, q \otimes q) = I(w, q \otimes p) + I(q \otimes p, q \otimes q) - \langle \ln q \otimes p - \ln q \otimes q, q \otimes p - w \rangle$$

Observe that $I(q \otimes p, q \otimes q) = I(p, q)$:

$$\langle \ln q \otimes p - \ln q \otimes q, q \otimes p \rangle = \langle 1_A \otimes (\ln p - \ln q), q \otimes p \rangle = \langle 1_A, q \rangle \langle \ln p - \ln q, p \rangle$$

and $\langle \ln q \otimes p - \ln q \otimes q, q \otimes p - w \rangle = 0$ by Lemma 1. $\qquad\square$

The relation between Shannon information $I(w, q \otimes p)$ and divergence $I(p, q)$ is illustrated on the diagram below:

We refer to $I(w, q \otimes q)$ as the *hypotenuse* of the channel or of the Markov morphism $T : \mathcal{P}(A) \to \mathcal{P}(B)$. This quantity allows us to relate information constraints in information-theoretic variational problems of the first and the third kind [11]. These problems are expressed by the following optimal value functions:

$$\overline{u}(\lambda) = \sup\{\langle u, p \rangle : I(p, q) \leq \lambda\}, \qquad\qquad p, q \in \mathcal{P}(B)$$
$$\overline{v}(\gamma) = \sup\{\langle v, w \rangle : I(w, q \otimes p) \leq \gamma\}, \qquad w, q \otimes p \in \mathcal{P}(A \otimes B)$$

It follows from the Shannon-Pythagorean theorem that the hypotenuse satisfies the constraint $I(w, q \otimes q) \leq \gamma + \lambda$. Thus, if u is the restriction $\langle v \rangle_A = u$ of operator $v \in A \otimes B$ to B, then both problems can be solved by maximizing $\langle v, w \rangle$ subject to constraint $I(w, q \otimes q) \leq \gamma + \lambda$. This property has been considered in problems of optimization of mutation rate control function in genetic algorithms [12].

References

1. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1) (1951) 79–86
2. Chentsov, N.N.: A nonsymmetric distance between probability distributions, entropy and the Pythagorean theorem. Mathematical notes of the Academy of Sciences of the USSR **4**(3) (September 1968) 323–332
3. Chentsov, N.N.: Statistical Decision Rules and Optimal Inference. Nauka, Moscow, U.S.S.R. (1972) In Russian, English translation: Providence, RI: AMS, 1982.
4. Umegaki, H.: Conditional expectation in an operator algebra. IV. entropy and information. Kodai Mathematical Seminar Reports **14**(2) (1962) 59–85
5. Araki, H.: Relative entropy of states of von Neumann algebras. Publications of the Research Institute for Mathematical Sciences **11**(3) (1975) 809–833
6. Belavkin, V.P., Staszewski, P.: Relative entropy in C∗-algebraic statistical mechanics. Reports in Mathematical Physics **20** (1984) 373–384
7. Khinchin, A.I.: Mathematical Foundations of Information Theory. Dover, New York (1957)
8. Rockafellar, R.T.: Conjugate Duality and Optimization. Volume 16 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, PA (1974)
9. Belavkin, R.V.: Utility and value of information in cognitive science, biology and quantum theory. In Accardi, L., Freudenberg, W., Ohya, M., eds.: Quantum Bio-Informatics III. Volume 26 of QP-PQ: Quantum Probability and White Noise Analysis. World Scientific (2010)
10. Ohya, M., Petz, D.: Quantum Entropy and its Use. Springer-Verlag, Berlin (1993)
11. Stratonovich, R.L.: Information Theory. Sovetskoe Radio, Moscow, USSR (1975) In Russian.
12. Belavkin, R.V.: Minimum of information distance criterion for optimal control of mutation rate in evolutionary systems. In Accardi, L., Freudenberg, W., Ohya, M., eds.: Quantum Bio-Informatics V. Volume 30 of QP-PQ: Quantum Probability and White Noise Analysis. World Scientific (2013)