RAJENDRA BHATIA AND JOHN HOLBROOK

Noncommutative Geometric Means

> For, in fact, what is man in nature? A Nothing in comparison with the Infinite, an All in comparison with the Nothing, a mean between nothing and everything.
>
> —*Blaise Pascal*

Averaging operations entered mathematics rather early. Fascinated as they were by geometric proportions, the ancient Greeks defined as many as eleven different means. The arithmetic, geometric, and harmonic means are the three best-known ones. If Pascal had one of these in mind when he composed his *Pensées* [P], he would soon have realised that mixing zero and infinity is a source of as many problems as mixing mathematics and divinity.

For centuries, mathematicians performed their operations either on numbers or on geometrical figures. Then in 1855 Arthur Cayley introduced new objects called *matrices*, and soon afterwards he gave the laws of their algebra. Seventy years later, Werner Heisenberg found that the noncommutativity of matrix multiplication offers just the right conceptual framework for describing the laws of atomic mechanics. Matrices were found to be useful in the description of classical vibrating systems and electrical networks as well. For mathematicians, analysis of linear operators was a subject of intense study throughout the twentieth century and into the twenty-first century.

Many quantities of basic interest such as states of quantum mechanical systems and impedances of electrical networks are defined in terms of matrices. Mixing of the underlying systems in various ways leads to corresponding operations on the matrices representing the systems. Not surprisingly, some of these are averaging operations or means.

Of the three most familiar means, the geometric mean combines the operations of multiplication and square roots. When we replace positive numbers by positive definite matrices, both of these operations involve new subtleties. In this article we introduce the reader to some of them.

◇ ◇ ◇

Let $\mathbb{R}_+$ be the set of all positive real numbers. Given a and b in $\mathbb{R}_+$ a *mean* $m(a,b)$ could be defined in different ways. It is reasonable to expect that the binary operation m on $\mathbb{R}_+$ has the following properties:

(i) $m(a,b) = m(b,a)$.
(ii) $\min(a,b) \leq m(a,b) \leq \max(a,b)$.
(iii) $m(\alpha a,\alpha b) = \alpha m(a,b)$ for all $\alpha > 0$.
(iv) $m(a,b)$ is an increasing function of a and b.
(v) $m(a,b)$ is a continuous function of a and b.

The three familiar means, arithmetic, geometric, and harmonic, satisfy all these requirements. Other examples of means include the *binomial means*, also called the *power means*, defined as

$$m_p(a,b) = \left(\frac{a^p + b^p}{2}\right)^{1/p}, \quad -\infty \leq p \leq \infty.$$

Here, it is understood that for the special values $p = 0$ and $\pm\infty$ we define $m_p(a,b)$ as the limits

$$\begin{aligned} m_0(a,b) &= \lim_{p \to 0} m_p(a,b) = \sqrt{ab}, \\ m_\infty(a,b) &= \lim_{p \to \infty} m_p(a,b) = \max(a,b), \\ m_{-\infty}(a,b) &= \lim_{p \to -\infty} m_p(a,b) = \min(a,b). \end{aligned}$$

The arithmetic and the harmonic means correspond to the cases $p = 1$ and -1, respectively. Inequalities between means have been studied for a long time. See the classic [HLP], and the more recent [BMV]. A sample result here is that for fixed a and b, $m_p(a,b)$ is an increasing function of p. This includes, as a special case, the inequality between the three familiar means.

There exists a fairly well-developed theory of means for positive definite matrices. Let $\mathbb{M}_n(\mathbb{C})$ be the set of all $n \times$

n complex matrices, $\mathbb{S}_n$ the collection of all self-adjoint elements of $\mathbb{M}_n(\mathbb{C})$, and $\mathbb{P}_n$ that of all positive definite matrices. The space $\mathbb{S}_n$ is a real vector space and $\mathbb{P}_n$ is an open cone within it. This gives rise to a natural order on $\mathbb{S}_n$. We say that $A \geq B$ if $A - B$ is positive definite or positive semidefinite. Two elements of $\mathbb{S}_n$ are not always comparable in this order. Every element X of GL_n (the group of invertible matrices) has a natural action on $\mathbb{P}_n$. This is given by the map $\Gamma_X(A) = X^*AX$. We say that A and B are *congruent* if $B = \Gamma_X(A)$ for some $X \in GL_n$. In the special case when X is unitary, we say that A and B are *unitarily equivalent*. The group of unitary matrices is denoted by $\mathbb{U}_n$.

Now we have enough structure to lay down conditions that a mean $M(A,B)$ of two positive definite matrices A and B should satisfy. Imitate the properties (i)–(v) for means of numbers. This suggests the following natural conditions:

(I) $M(A,B) = M(B,A)$.
(II) If $A \leq B$, then $A \leq M(A,B) \leq B$.
(III) $M(X^*AX,X^*BX) = X^*M(A,B)X$, for all $X \in GL_n$.
(IV) $M(A,B)$ is an increasing function of A and B; i.e., if $A_1 \geq A_2$ and $B_1 \geq B_2$, then $M(A_1,B_1) \geq M(A_2,B_2)$.
(V) $M(A,B)$ is a continuous function of A and B.

The monotonicity condition (IV) is a source of many intriguing problems in constructing matrix means. This is because the order $A \geq B$ is somewhat subtle. For example, if

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

then $A \geq B$ but $A^2 \ngeq B^2$.

What functions of positive numbers, when lifted to positive definite matrices, preserve order? This is the subject of an elegant and richly applicable theory developed by Charles Loewner. Let f be a real-valued function on $\mathbb{R}_+$. If A is a positive definite matrix and $A = \Sigma\lambda_i u_i u_i^*$ is its spectral resolution, then $f(A)$ is the self-adjoint matrix defined as $f(A) = \Sigma f(\lambda_i)u_i u_i^*$. We say that f is a *matrix monotone function* if for all $n = 1, 2, \ldots$, the inequality $A \geq B$ in $\mathbb{P}_n$ implies $f(A) \geq f(B)$. One of the theorems of Loewner says that f is matrix monotone if and only if it has an analytic continuation to a mapping of the upper half-plane into itself. As a consequence, the function $f(x) = x^p$ is matrix monotone if and only if $0 \leq p \leq 1$. The function $f(x) = \log x$ is matrix monotone, but $f(x) = \exp x$ is not. We refer the reader to Chapter V of [B] for an exposition of Loewner's theory.

Returning to means, the arithmetic and the harmonic means of A and B are defined, in the obvious way, as $\frac{1}{2}(A + B)$ and $[\frac{1}{2}(A^{-1} + B^{-1})]^{-1}$, respectively. It is easy to see that they satisfy the conditions (I)–(V) above.

The notion of geometric mean in this context is more elusive, even treacherous. Every positive definite matrix A has a unique positive definite square root $A^{1/2}$. However, if A and B are positive definite, then unless A and B commute, the product $A^{1/2}B^{1/2}$ is not self-adjoint, let alone positive definite. This rules out using $A^{1/2}B^{1/2}$ as our geometric mean of A and B, except in the trivial case when $AB = BA$. We should look for other good expressions in A and B that reduce to $A^{1/2}B^{1/2}$ when A and B commute. One plausible choice is the quantity

$$\text{(1)} \quad \exp\left(\frac{\log A + \log B}{2}\right) = \lim_{p \to 0}\left(\frac{A^p + B^p}{2}\right)^{1/p}.$$

The equality of the two sides of (1) was noted by Bhagwat and Subramanian [BS], who studied in detail the "power means" occurring on the right-hand side. This too is not monotone in A and B, as can be seen by choosing positive definite matrices X and Y, for which $X \geq Y$ but $\exp X \ngeq \exp Y$, and then choosing A and B such that $X = \frac{1}{2}(\log A + \log B)$ and $Y = \frac{1}{2}\log B$.

The condition (III), sometimes called the *transformer equation*, is not innocuous either. Our failed candidates fail on this count too.

The noncommutative analogue of $\sqrt{ab}$ with all desirable properties turns out to be the expression

$$\text{(2)} \qquad A\#B = A^{1/2}\,(A^{-1/2}\,BA^{-1/2})^{1/2}\,A^{1/2},$$

that was introduced by Pusz and Woronowicz [PW] in 1975. At the outset it does not *appear* to be symmetric in A and B; but it is, as we will soon see. The monotonicity in B is assured by the facts that congruence preserves order ($B_1 \geq B_2$ implies $X^*B_1X \geq X^*B_2X$) and the square root function is matrix monotone.

Symmetry in A and B is apparent more easily from an alternative characterisation of $A\#B$ due to T. Ando [A]. We have

$$\text{(3)} \qquad A\#B = \max\left\{X : X = X^* \text{ and } \begin{bmatrix} A & X \\ X & B \end{bmatrix} \geq 0\right\}.$$

Among its other characterisations, one describes $A\#B$ as the unique positive definite solution of the Riccati equation

$$\text{(4)} \qquad XA^{-1}X = B.$$

We call $A\#B$ the *geometric mean* of A and B. It has the desired properties (I)–(V) expected of a mean $M(A,B)$: property (III) may be verified easily from (3) or (4). It satisfies the expected inequality

$$\text{(5)} \qquad \left(\frac{A^{-1} + B^{-1}}{2}\right)^{-1} \leq A\#B \leq \frac{A + B}{2},$$

and has other pleasing properties. Many of these were derived by Ando [A].

Two positive definite matrices A and B can be diagonalised simultaneously by a unitary conjugation Γ_U if and only if they commute. In the absence of commutativity, A and B can be diagonalised simultaneously by a congruence in two steps:

$$(A,B) \xrightarrow{\Gamma_{A^{-1/2}}} (I,A^{-1/2}\,BA^{-1/2}) \xrightarrow{\Gamma_U} (I,D),$$

where U is a unitary such that $U^*\,(A^{-1/2}BA^{-1/2})U$ is a diagonal matrix D. This takes some of the mystery out of the formula (2). In fact, any mean $m(a,b)$ of positive numbers leads to a mean $M(A,B)$ of positive definite matrices by the procedure $M(A,B) = \Gamma_{A^{1/2}}(m(I,D))$. To ensure that M is an increasing function of A and B, we have to assume that the function $f(x) = m(1,x)$ is matrix monotone. The formula (2) corresponds to the case when $m(a,b) = (ab)^{1/2}$.

The indirect argument we have used to deduce the symmetry of the geometric mean is not necessary. Let $m(a,b)$ be any mean, let $f(x) = m(1,x)$, and

$$M(A,B) = A^{1/2} f\,(A^{-1/2}BA^{-1/2})\, A^{1/2}. \tag{6}$$

Though this expression seems to be asymmetric in A and B, in fact $M(A,B) = M(B,A)$. For this we need to prove

$$f(A^{-1/2}BA^{-1/2}) = A^{-1/2}B^{1/2}f\,(B^{-1/2}AB^{-1/2})B^{1/2}A^{-1/2}.$$

Using the polar decomposition $A^{-1/2}B^{1/2} = PU$, where P is positive definite and U unitary, this statement reduces to

$$f(P^2) = PU\,f(U^*P^{-2}U)U^*P = P\,f(P^{-2})P.$$

This, in turn, is equivalent to saying that for every eigenvalue λ of P, we have

$$m(1,\lambda^2) = \lambda m(1,\lambda^{-2})\lambda.$$

But that is a consequence of properties (i) and (iii) of the mean m. A similar argument verifies (III).

A simple corollary of this construction is the persistence of inequalities like (5) when one passes from positive numbers to positive definite matrices. Kubo and Ando [KA] developed a general theory of matrix means and established a correspondence between such means and matrix monotone functions.

What happens when we have three positive definite matrices instead of two? The arithmetic and the harmonic means present no problems. Plainly, they should be defined as $\frac{1}{3}(A + B + C)$ and $[\frac{1}{3}(A^{-1} + B^{-1} + C^{-1})]^{-1}$, respectively. The geometric mean, once again, raises interesting problems.

We would like to have a geometric mean $G(A,B,C)$ that reduces to $A^{1/3}B^{1/3}C^{1/3}$ when A, B, and C commute with each other. In addition it should have the following properties.

(α) $G(A,B,C) = G(\pi(A,B,C))$ for any permutation π of the triple (A,B,C).
(β) $G(X^*AX,X^*BX,X^*CX) = X^*G(A,B,C)X$ for all $X \in GL_n$.
(γ) $G(A,B,C)$ is an increasing function of A, B, and C.
(δ) $G(A,B,C)$ is a continuous function of A, B, and C.

None of the procedures presented above for two matrices extends readily to three. The expressions (2), (3), and (4) have no obvious generalisations that work. The idea of simultaneous diagonalisation does not help either: while two positive definite matrices can be diagonalised simultaneously by a congruence, generally three can not be. Defining a suitable geometric mean of three positive definite matrices has been a ticklish problem for many years. Recently some progress has been made in this direction, and we describe it now.

◇ ◇ ◇

One geometry cannot be more true than another; it can only be more convenient.

—Henri Poincaré [Po]

While the geometric mean $A\#B$ has been much studied in connection with problems of matrix analysis, mathematical physics, and electrical engineering, a deeper understanding of it is achieved by linking it with some standard constructions in Riemannian geometry.

The space $\mathbb{M}_n(\mathbb{C})$ has a natural inner product $\langle A,B\rangle = \operatorname{tr} A^*B$. The associated norm $\|A\|_2 = (\operatorname{tr} A^*A)^{1/2}$ is called the Frobenius, or the Hilbert-Schmidt, norm. If A is a matrix with eigenvalues $\lambda_1, \dots, \lambda_n$, we write $\lambda(A)$ for the vector $(\lambda_1, \dots, \lambda_n)$ or for the diagonal matrix $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$.

The set $\mathbb{P}_n$ is an open subset of $\mathbb{S}_n$ and thus is a differentiable manifold. The exponential is a bijection from $\mathbb{S}_n$ onto $\mathbb{P}_n$. The Riemannian metric on the manifold $\mathbb{P}_n$ is constructed as follows. The element of arc length is the differential

$$ds = \|A^{-1/2}\, dA\; A^{-1/2}\|_2. \tag{7}$$

This gives the prescription for computing the length of a differentiable curve in $\mathbb{P}_n$. If $\gamma : [a,b] \to \mathbb{P}_n$ is such a curve, then its length, obtained by integrating the formula (7), is

$$L(\gamma) = \int_a^b \|\gamma^{-1/2}(t)\gamma'(t)\gamma^{-1/2}(t)\|_2\, dt. \tag{8}$$

If A and B are two elements of $\mathbb{P}_n$, then among all curves γ joining A and B there is a unique one of minimum length. This is called the *geodesic* joining A and B. We write this curve as $[A,B]$, and denote its length, as defined by (8), by the symbol $\delta_2(A,B)$. This gives a metric on $\mathbb{P}_n$ called the Riemannian metric.

From the invariance of trace under similarities, it is easy to see that for every X in GL_n the map $\Gamma_X : \mathbb{P}_n \to \mathbb{P}_n$ is a bijective isometry on the metric space $(\mathbb{P}_n,\delta_2)$.

An important feature of this metric is the *exponential metric increasing property* (EMI). This says that the map exp from the metric space $(\mathbb{S}_n,\|\cdot\|_2)$ to $(\mathbb{P}_n,\delta_2)$ increases distances. More precisely, if H and K are Hermitian matrices, then

$$\|H - K\|_2 \le \delta_2(e^H,e^K). \tag{9}$$

To prove this, one uses the formula (8) and an infinitesimal version of (9):

$$\|K\|_2 \le \|e^{-H/2}\, De^H(K)e^{-H/2}\|_2 \tag{10}$$

for all $H, K \in \mathbb{S}_n$. Here $De^H(K)$ is the derivative of the map exp at the point H evaluated at K, i.e.,

$$De^H(K) = \lim\nolimits_{t\to 0} \frac{e^{H+tK} - e^H}{t}. \tag{11}$$

There is a well-known formula due to Daleckii and Krein (see [B], chapter V, for example) giving an expression for this derivative. Choose an orthonormal basis in which $H = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Then

$$De^H(K) = \left[\frac{e^{\lambda_i} - e^{\lambda_j}}{\lambda_i - \lambda_j} k_{ij}\right].$$

(The notation here is that $[x_{ij}]$ stands for a matrix with entries x_{ij}.) From this, one sees that the (i,j) entry of $e^{-H/2}De^H(K)e^{-H/2}$ is

$$\frac{\sinh(\lambda_i - \lambda_j)/2}{(\lambda_i - \lambda_j)/2} k_{ij}. \tag{12}$$

Since $\frac{\sinh x}{x} \ge 1$, the inequality (10) follows from this.

In the special case when H and K commute, a calculation shows that there is equality in (9). In this case the function exp maps the line segment $[H,K]$ in the Euclidean space $\mathbb{S}_n$ isometrically onto the geodesic segment $[e^H,e^K]$ in $\mathbb{P}_n$. If $A = e^H$ and $B = e^K$, this says that the geodesic segment joining A and B is the path

$$\gamma(t) = e^{(1-t)H+tK} = e^{(1-t)H}e^{tK} = A^{1-t}\,B^t, \quad 0 \leq t \leq 1.$$

Further, $\delta_2(A,\gamma(t)) = t\delta_2(A,B)$ for each t in [0,1].

The case of noncommuting A and B can be reduced to the commuting case using the fact that $\Gamma_{A^{-1/2}}$ is an isometry on the space $(\mathbb{P}_n,\delta_2)$. The geodesic segment $[I,A^{-1/2}\,BA^{-1/2}]$ is parametrised by $\gamma_0(t) = (A^{-1/2}BA^{-1/2})^t$, by what we said about the commuting case. So, the geodesic $[A,B] = [\Gamma_{A^{1/2}}(I),\Gamma_{A^{1/2}}(A^{-1/2}BA^{-1/2})]$ is parametrised by

$$(13) \qquad \gamma(t) = A^{1/2}\,(A^{-1/2}\,BA^{-1/2})^t\,A^{1/2},\ 0 \leq t \leq 1.$$

This shows that the geometric mean $A\#B$ defined by the formula (2) is nothing but the midpoint of the geodesic joining A and B in the Riemannian manifold $\mathbb{P}_n$. Thus while (2), (3), and (4) might have appeared as over-imaginative noncommutative variants of $\sqrt{ab}$, very natural geometric considerations lead to the same notion of *mean* as is given by (2). Note that for each t, $\gamma(t)$ defined by (13) is a mean of A and B corresponding to the function $f(x) = x^t$ in the formula (6). Those means are not symmetric, however: (I) fails unless $t = 1/2$.

This discussion also gives an explicit formula for the metric δ_2. We have $\delta_2(A,B) = \delta_2\ (I,A^{-1/2}BA^{-1/2}) = \|\log I - \log\,(A^{-1/2}BA^{-1/2})\|_2 = \|\log(A^{-1/2}BA^{-1/2})\|_2$. The matrices $A^{-1/2}BA^{-1/2}$ and $A^{-1}B$ have the same eigenvalues. So, this can be expressed as

$$(14) \quad \delta_2(A,B) = \|\log \lambda(A^{-1}B)\|_2 = \left(\sum_{i=1}^{n}(\log \lambda_i(A^{-1}B))^2\right)^{1/2}.$$

The inequality (9) captures an essential feature of $\mathbb{P}_n$: it is a manifold of nonpositive curvature. To understand this, consider a triangle with three vertices O, H, and K in $\mathbb{S}_n$. Under the exponential map, this is mapped to a "triangle" with vertices I, exp H and exp K in $\mathbb{P}_n$. The lengths of the two sides $[O,H]$ and $[O,K]$ measured by the norm $\|\cdot\|_2$ are equal to the lengths of their images $[I$, exp $H]$ and $[I$, exp $K]$ measured by the metric δ_2. By the EMI (9), the length of the third side [exp H, exp K] of the triangle in $\mathbb{P}_n$ is larger than (or equal to) $\|H - K\|_2$. The general case of a geodesic triangle with vertices exp A, exp B, exp C in $\mathbb{P}_n$ may be reduced to the special case by applying the congruence $\Gamma_{\exp(-A/2)}$ to all points and thus changing one of the vertices to I. This is often described by saying that two geodesics emanating from a point in $\mathbb{P}_n$ spread out faster than their pre-images (under the exponential map) in $\mathbb{S}_n$.

It is instructive here to compare the situation with that of $\mathbb{U}_n$, a compact manifold of non-negative curvature (Figure 1). In this case the real vector space $i\mathbb{S}_n$ consisting of skew-Hermitian matrices is mapped by the exponential onto $\mathbb{U}_n$. The map is not injective; it is a local diffeomorphism.

Using the formula (11) with H and K in $i\mathbb{S}_n$, we reduce

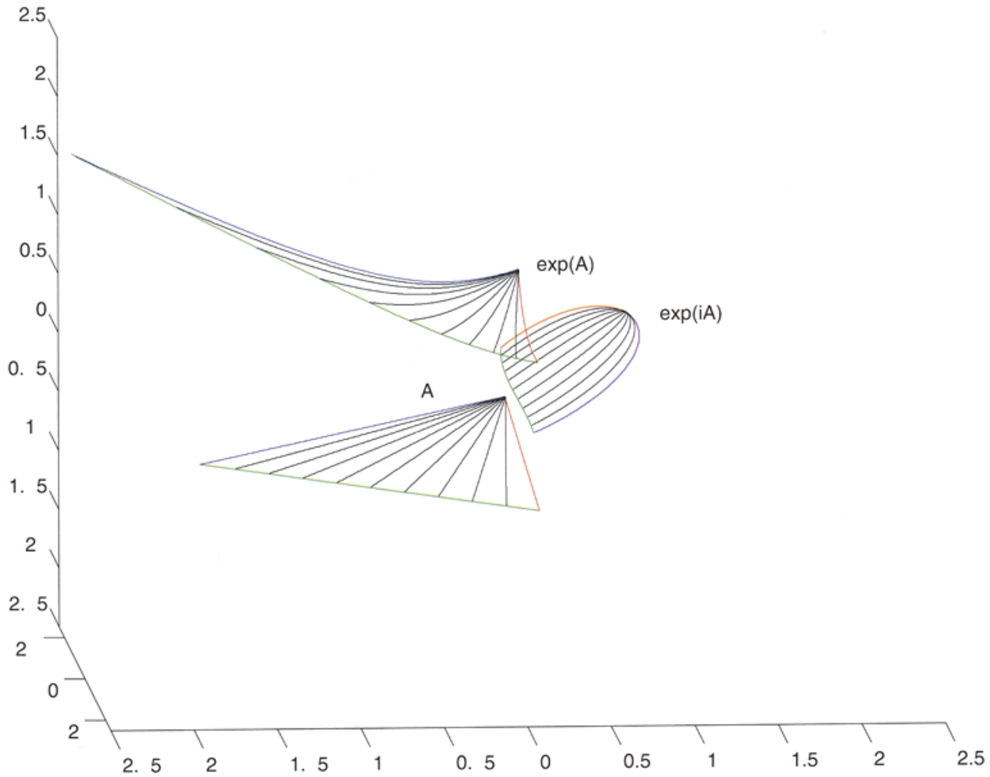


Figure 1. Three curvatures, showing a comparison of a Euclidean (curvature zero) triangle in $\mathbb{S}_2$ with its images under exp(·) in $\mathbb{P}_2$ (nonpositive curvature) and exp(*i*·) in $\mathbb{U}_2$ (non-negative curvature). The colours indicate matching vertices. Note that the geodesics emanating from exp(*A*) spread out faster than Euclidean ones (compare the straight lines at *A*), whereas those emanating from exp(*iA*) spread more slowly.

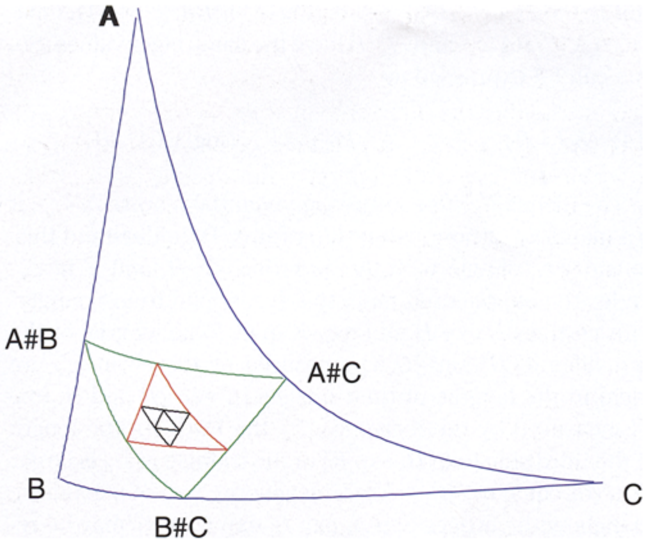


Figure 2. Geodesic distance from *A#B* to *A#C* is no more than half that from *B* to *C*. Joining the midpoints of the sides of a geodesic triangle in $\mathbb{P}_n$ results in a triangle with sides no more than half as long. Iterating this procedure leads to the construction of Ando, Li, and Mathias, described in the text.

H to diag$(i\lambda_1, \ldots, i\lambda_n)$ with λ_j real. Instead of (12) we have now

$$\frac{\sin(\lambda_i - \lambda_j)/2}{(\lambda_i - \lambda_j)/2} k_{ij}.$$

Since $\left|\frac{\sin x}{x}\right| \le 1$, the inequality (10) is reversed in this case, as is its consequence (9), provided e^H and e^K are close to each other.

Returning to $\mathbb{P}_n$ and the geometric mean, it is not difficult to derive from the information at our disposal the fact that given any three points A, B, and C in $\mathbb{P}_n$ we have

$$(15) \qquad \delta_2(A\#B,A\#C) \le \frac{1}{2}\delta_2(B,C).$$

This inequality says that in every geodesic triangle in $\mathbb{P}_n$ with vertices A, B, and C, the length of the geodesic joining the midpoints of two sides is at most half the length of the third side. (If the geometry were Euclidean, the two sides of (15) would have been equal.) Figure 2 illustrates (15).

We saw that the geometric mean $A\#B$ is the midpoint of the geodesic $[A,B]$. This suggests that we may possibly define the geometric mean of three positive definite matrices A, B, and C as the "centroid" of the geodesic triangle $\Delta(A,B,C)$ in $\mathbb{P}_n$.

In a Euclidean space $\mathscr{E}$, the centroid $\bar{x}$ of a triangle with vertices x_1, x_2, x_3 is the point $\bar{x} = \frac{1}{3}(x_1 + x_2 + x_3)$. This is the arithmetic mean of the vectors x_1, x_2, and x_3. This point may be characterised by several other properties. Three of them are:

(M1) $\bar{x}$ is the unique point of intersection of the three medians of the triangle $\Delta(x_1,x_2,x_3)$, as in Figure 3;

(M2) $\bar{x}$ is the unique point in $\mathscr{E}$ at which the function

$$\|x - x_1\|^2 + \|x - x_2\|^2 + \|x - x_3\|^2$$

attains its minimum;

(M3) $\bar{x}$ is the unique point of intersection of the nested sequence of triangles $\{\Delta_n\}$ in which $\Delta_1 = \Delta$ and Δ_{j+1} is the triangle obtained by joining the mid-

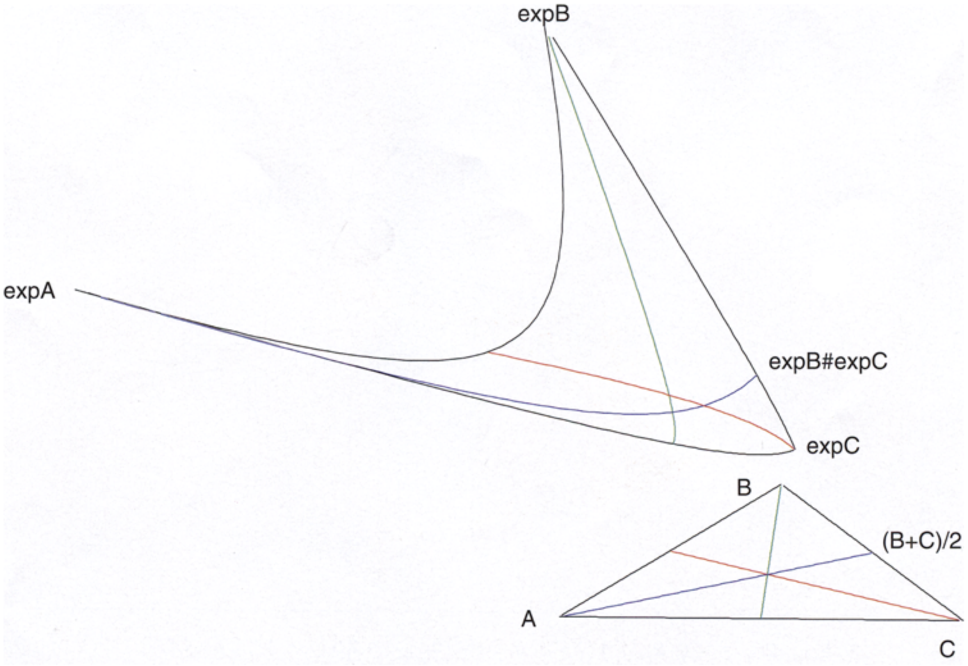


Figure 3. In the hyperbolic geometry medians may not meet. While the medians of a Euclidean triangle intersect at the centroid, the corresponding median geodesics of a triangle in $\mathbb{P}_n$ may not intersect at all. A 3-D wire model would make it clear that, generically, the medians do not even intersect in pairs.

points of the three sides of Δ_j (Figure 2 mimics this construction in the non-Euclidean setting of $\mathbb{P}_n$).

To define a geometric mean of A, B, and C in $\mathbb{P}_n$ we may try to imitate one of these definitions, now modified to suit the geometry of $\mathbb{P}_n$. Here fundamental differences between Euclidean and hyperbolic geometry come to the fore, and (M1), (M2), and (M3) lead to three different results.

The first definition using (M1) fails. The triangle $\Delta(A,B,C)$ may be defined as the "convex set" generated by A, B, and C. (It is clear what that should mean: replace line segments in the definition of convexity by geodesic segments.) It turns out that this is not a 2-dimensional object as in ordinary Euclidean geometry (see Figure 4). So, the medians of a triangle may not intersect at all in some cases (again, see Figure 3).

With (M2) as our motivation, we may ask whether there exists a point X_0 in $\mathbb{P}_n$ at which the function

$$f(X) = \delta_2^2(A,X) + \delta_2^2(B,X) + \delta_2^2(C,X)$$

attains a minimum. It was shown by Élie Cartan (see, for example, section 6.1.5 of [Be]) that given A, B, and C in $\mathbb{P}_n$, there is a unique point X_0 at which f has a minimum. Let $G_2(A,B,C) = X_0$, and think of it as a geometric mean of A, B, and C. This mean has been studied in two recent papers by Bhatia and Holbrook [BH] and Moakher [M].

In another recent paper [ALM], Ando, Li, and Mathias define a geometric mean $G_3(A,B,C)$ by an iterative procedure. This iterative procedure has a nice geometric interpretation: it amounts to reaching the centroid of the geodesic triangle $\Delta(A,B,C)$ in $\mathbb{P}_n$ by a process akin to (M3). Starting with Δ_1 as the triangle $\Delta(A,B,C)$ one defines Δ_2 to be $\Delta(A\#B,A\#C,B\#C)$, and then iterates this process. Figure 2 shows the beginning of this process. The inequality (15) guarantees that the diameters of these nested triangles descend to zero as $1/2^n$. It can then be seen that there is a unique point in the intersection of this decreasing sequence of triangles. This point, represented by $G_3(A,B,C)$, is the geometric mean proposed by Ando, Li, and Mathias.

It turns out that the two objects $G_2(A,B,C)$ and $G_3(A,B,C)$ are not always equal (Figure 5 illustrates this phenomenon). Thus we have (at least) two competing notions of the centroid of $\Delta(A,B,C)$. How do they do as geometric means? The mean $G_3(A,B,C)$ has all of the four desirable properties (α)–(δ) that we listed for a mean $G(A,B,C)$. Properties (α), (β), and (δ) are almost obvious from the construction. Property (γ)—monotonicity—is a consequence of the fact that the geometric mean $A\#B$ is monotone in A and B. So monotonicity is preserved at each iteration step. The mean $G_2(A,B,C)$ does have the desirable properties (α), (β), and (δ). Property (β) follows from the fact that Γ_X is an isometry of $(\mathbb{P}_n,\delta_2)$ for every X in GL_n. However, we have not been able to prove that $G_2(A,B,C)$ is monotone in A, B, and C. We have an unresolved question: Given positive definite matrices A, B, C, and A' with $A \geq A'$, is $G_2(A,B,C) \geq G_2(A',B,C)$?

An answer to this question may lead to better understanding of the geometry of $\mathbb{P}_n$, the best-known example of a manifold of nonpositive curvature. Certainly this is of interest in matrix analysis. Computer experiments suggest an affirmative answer to the question.

Finally, we make a brief mention of two related matters. The Frobenius norm is one of a large class of norms called

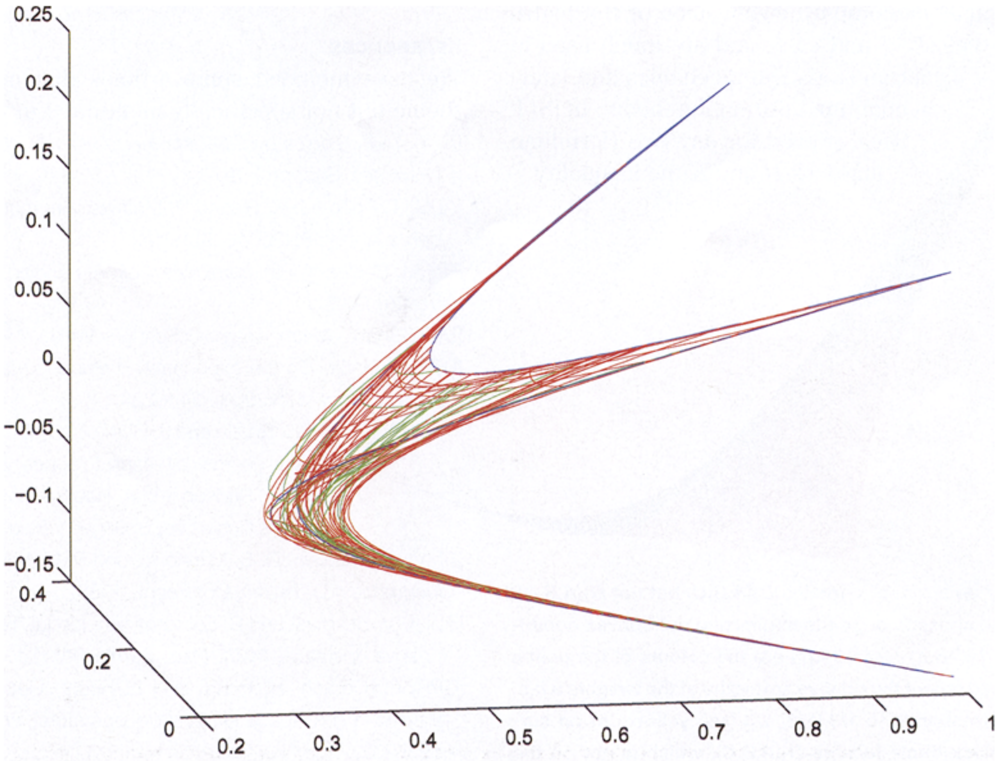


Figure 4. Conv (A,B,C) is not two-dimensional. In the hyperbolic (nonpositive curvature) geometry of $\mathbb{P}_n$, the convex hull of a triangle (formed by successively adjoining the geodesics between points that are already in the object) is not a surface but rather a "fatter" object.

unitarily invariant norms or *Schatten-von Neumann norms*. These norms $\|\cdot\|_\Phi$ have the invariance property $\|UAV\|_\Phi = \|A\|_\Phi$ for all unitary U and V. Each of these norms corresponds to a *symmetric norm* Φ on $\mathbb{R}^n$; that is, a norm Φ that is invariant under permutations and sign changes of coordinates. The correspondence is given by $\|A\|_\Phi = \Phi\,(s_1(A), \ldots, s_n(A))$, where $s_1(A) \geq \cdots \geq s_n(A)$ are the singular values of A. Common examples are the *Hölder norms* $\Phi_p(x) = (\Sigma|x_j|^p)^{1/p}$ and the corresponding *Schatten norms* $\|A\|_p = (\Sigma\, s_j^p(A))^{1/p}$, $1 \leq p \leq \infty$. The Frobenius norm is the special case $p = 2$.

For each of these norms we may define a metric δ_Φ on $\mathbb{P}_n$ as in the formula (14). The EMI in the form (9) or (10) remains true (see [B2]). The import of this remark is that, with any of these metrics, $\mathbb{P}_n$ is a *Finsler manifold* of non-positive curvature; the special Frobenius norm arises from an inner product and gives rise to a Riemannian structure. In recent years *metric spaces of nonpositive curvature* have been studied in great detail; see the comprehensive book by Bridson and Haefliger [BrHa]. The spaces $\mathbb{P}_n$ with norms $\|\cdot\|_\Phi$ are interesting and natural examples of such spaces.

◇ ◇ ◇

> But the whole wondrous complications of interference, waves, and all, result from the little fact that $\hat{x}\hat{p} - \hat{p}\hat{x}$ is not quite zero.
>
> —*Richard Feynman [FLS]*

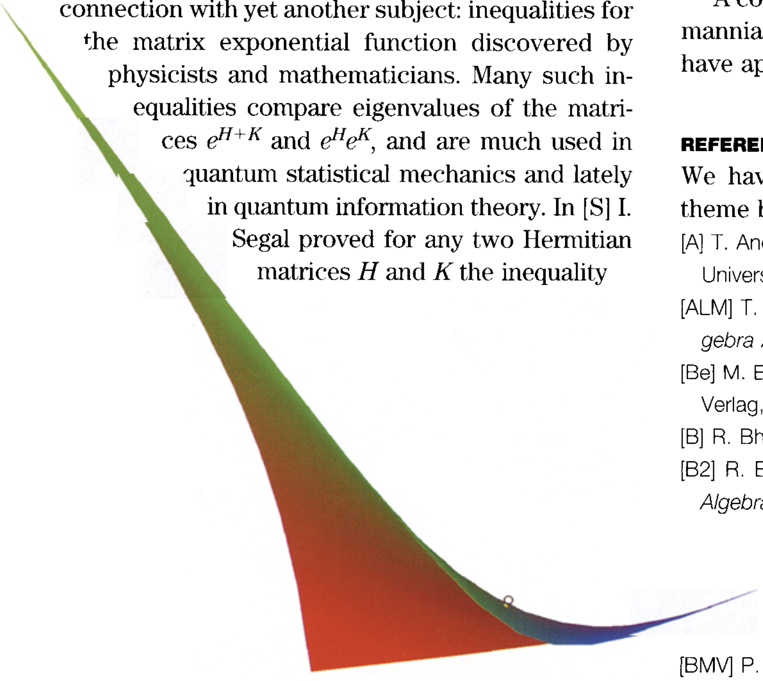


Figure 5. The "Cartan surface" contains $G_2(A,B,C)$ but not $G_3(A,B,C)$. The Cartan surface consists of points minimizing the convex combinations $a\delta_2^2(A,X) + b\delta_2^2(B,X) + c\delta_2^2(C,X)$; here the colours of the points shown are chosen to reflect the relative strengths of the weights a,b,c. Thus $G_2(A,B,C)$ corresponds to 1/3, 1/3, 1/3 (see yellow dot on surface). The small black circle locates $G_3(A,B,C)$, which is not on the surface in general. Thanks to J.-P. Shoch for computing this picture of a Cartan surface.

The generalised version of EMI has a fascinating connection with yet another subject: inequalities for the matrix exponential function discovered by physicists and mathematicians. Many such inequalities compare eigenvalues of the matrices e^{H+K} and e^He^K, and are much used in quantum statistical mechanics and lately in quantum information theory. In [S] I. Segal proved for any two Hermitian matrices H and K the inequality

$$(16) \quad \lambda_1(e^{H+K}) \leq \lambda_1\,(e^{H/2}e^Ke^{H/2}).$$

Here $\lambda_1(X)$ is the largest eigenvalue of a matrix X with real eigenvalues. In a similar vein, we have the famous Golden-Thompson inequality

$$(17) \qquad \operatorname{tr}\,(e^{H+K}) \leq \operatorname{tr}\,(e^{H/2}e^Ke^{H/2}).$$

The matrices e^{H+K} and $e^{H/2}e^Ke^{H/2}$ are positive definite. So, the inequalities (16) and (17) say

$$\|e^{H+K}\|_p \leq \|e^{H/2}e^Ke^{H/2}\|_p, \text{ for } p = 1,\infty.$$

The EMI (9) generalised to all unitarily invariant norms is the inequality

$$(18) \qquad \|H + K\|_\Phi \leq \|\log\,(e^{H/2}e^Ke^{H/2})\|_\Phi.$$

By well-known properties of the matrix exponential, this implies

$$(19) \qquad \|e^{H+K}\|_\Phi \leq \|e^{H/2}e^Ke^{H/2}\|_\Phi.$$

This inequality, called the generalised Golden-Thompson inequality, includes in it the inequalities (16) and (17). The origins of these inequalities and their connections with quantum statistical mechanics are explained in Simon [Si] (page 94). Still more general versions have been discovered by Lieb and Thirring, and by Araki, again in connection with problems of quantum physics. See Chapter IX of [B]. Generalisations in a different direction were opened up by Kostant [K], where the matrix exponential is replaced by the exponential map in more abstract Lie groups.

A common thread running between matrix analysis, Riemannian and Finsler geometry, and physics! Pascal would have approved.

REFERENCES

We have included some articles that are related to our theme but not specifically mentioned in the text.

[A] T. Ando, *Topics on Operator Inequalities*, Lecture Notes, Hokkaido University, Sapporo, 1978.

[ALM] T. Ando, C.-K. Li, and R. Mathias, Geometric means, *Linear Algebra Appl.* 385(2004), 305–334.

[Be] M. Berger, *A Panoramic View of Riemannian Geometry*, Springer-Verlag, 2003.

[B] R. Bhatia, *Matrix Analysis*, Springer-Verlag, 1997.

[B2] R. Bhatia, On the exponential metric increasing property, *Linear Algebra Appl.* 375(2003), 211–220.

[BH] R. Bhatia and J. Holbrook, Riemannian geometry and matrix geometric means, to appear in *Linear Algebra Appl.*

[BrHa] M. Bridson and A. Haefliger, *Metric Spaces of Non-positive Curvature*, Springer-Verlag, 1999.

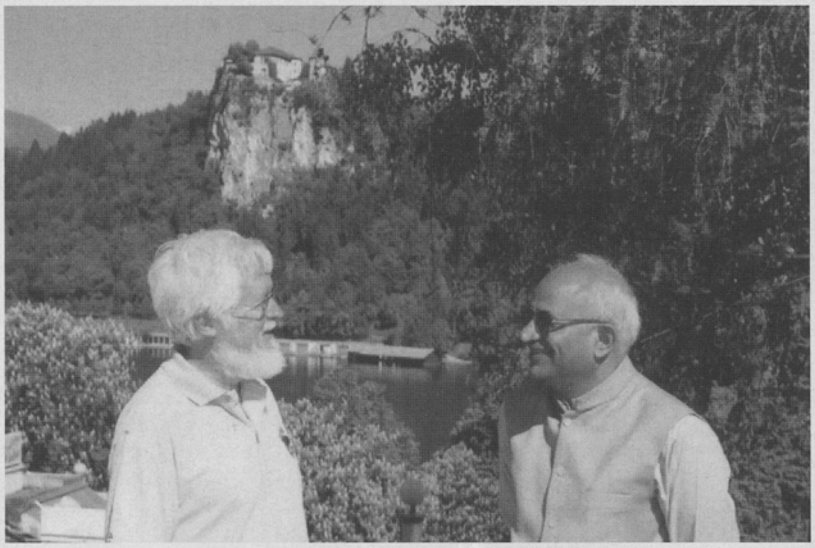
[BMV] P. S. Bullen, D. S. Mitrinovic, and P. M. Vasic, *Means and Their Inequalities*, D. Reidel, Dordrecht, 1988.

[BS] K. V. Bhagwat and R. Subramanian, Inequalities between means of positive operators, *Math. Proc. Camb. Phil. Soc.* 83(1978), 393–401.

[CPR] G. Corach, H. Porta, and L. Recht, Geodesics and operator means in the space of positive operators, *Int. J. Math.* 4(1993), 193–202.

[FLS] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, volume 3, page 20–17, Addison-Wesley, 1965.

AUTHORS



JOHN HOLBROOK
Department of Mathematics and Statistics
University of Guelph
Guelph, Ontario N1G 2W1
Canada
e-mail: jholbroo@uoguelph.ca

RAJENDRA BHATIA
Indian Statistical Institute
7, S. J. S. Sansanwal Marg
New Delhi 110016
India
e-mail: rbh@isid.ac.in

Rajendra Bhatia did his doctoral studies at ISI Delhi with Kalyan Mukherjee. He has been based there most of the quarter-century since, along with his wife Irpinder and their son Gautam.

John Holbrook is now Professor Emeritus at Guelph. He and his wife Catherine divide their time between Guelph and Fowke Lake (farther north), generally in the company of children, grandchildren, and cats.

This photograph of the authors (courtesy of Peter Šemrl) shows them in yet another continent, Europe: in the beautiful Alps of Slovenia. The photo may also serve as encouraging evidence that it is possible to collaborate on mathematical projects and remain on good terms!

[HLP] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, 1934.

[K] B. Kostant, On convexity, the Weyl group and the Iwasawa decomposition, *Ann. Sc. E. N. S.* 6(1973), 413–455.

[KA] F. Kubo and T. Ando, Means of positive linear operators, *Math. Ann.* 246(1980), 205–224.

[LL] J. D. Lawson and Y. Lim, The geometric mean, matrices, metrics, and more, *Amer. Math. Monthly* 108(2001), 797–812.

[M] M. Moakher, A differential geometric approach to the geometric mean of symmetric positive-definite matrices, *SIAM J. Matrix Anal. Appl.* 26(2005), 735–747.

[P] B. Pascal, *Pensées*, translation by W. F. Trotter, excerpt from item 72, Encyclopaedia Britannica, Great Books 33, 1952.

[Po] H. Poincaré, *Science and Hypothesis*, from page 50 of the Dover reprint, Dover Publications, 1952.

[PW] W. Pusz and S. L. Woronowicz, Functional calculus for sesquilinear forms and the purification map, *Reports Math. Phys.* 8(1975), 159–170.

[S] I. Segal, Notes towards the construction of nonlinear relativistic quantum fields III, *Bull. Amer. Math. Soc.* 75(1969), 1390–1395.

[Si] B. Simon, *Trace Ideals and Their Applications*, Cambridge University Press, 1979.