The Mathematics Student ISSN: 0025-5742
Vol. 88, Nos. 3-4, July-Dec (2019), xx–yy

THE CAUCHY-SCHWARZ INEQUALITY IN MATHEMATICS, PHYSICS AND STATISTICS

M. RAM MURTY

(Received: 22 - 03 - 2019; Revised: 07 - 05 - 2019)

Abstract. We discuss the Cauchy-Schwarz inequality, first in the mathematical setting, and then in physics formulated as Heisenberg's uncertainty principle in quantum mechanics and then in statistics manifesting as the Cramér-Rao inequality.

1. Cauchy-Schwarz inequality in mathematics

Perhaps the most ubiquitous of inequalities in mathematics is the Cauchy-Schwarz inequality. First discovered by Cauchy in the year 1821, it states that if $a_1, ..., a_n$ and $b_1, ..., b_n$ are arbitrary real numbers, then

$$\left|\sum_{j=1}^n a_j b_j\right| \leq \left(\sum_{j=1}^n |a_j|^2\right)^{1/2} \left(\sum_{j=1}^n |b_j|^2\right)^{1/2}, \tag{1.1}$$

with equality arising if and only if there is a $\lambda \in \mathbb{R}$ such that $a_j = \lambda b_j$ for $j = 1, 2, ..., n$.

There are several immediate proofs of this. Indeed, we have

$$\begin{aligned}\sum_{j=1}^n a_j^2 \sum_{k=1}^n b_k^2 - \left(\sum_{j=1}^n a_j b_j\right)^2 &= \sum_{j,k=1}^n a_j^2 b_k^2 \ - \sum_{j,k=1}^n a_j b_j a_k b_k \\ &= \sum_{j=1}^n \sum_{k\geq j} (a_j b_k - a_k b_j)^2 \geq 0. \end{aligned} \tag{1.2}$$

From this, we see that equality arises if and only if $a_j b_k = a_k b_j$ for all j, k, which is tantamount to the assertion

$$(a_1, ..., a_n) = \lambda(b_1, ..., b_n), \tag{1.3}$$

for some $\lambda \in \mathbb{R}$.

Another "proof at a glance" begins with the self-evident inequality

$$2a_j b_j \leq a_j^2 + b_j^2, \qquad 1 \leq j \leq n,$$

and noting the homogeneity of the left hand side, gets for any $\lambda \neq 0$,

$$2a_j b_j \leq \lambda^2 a_j^2 + \lambda^{-2} b_j^2, \qquad 1 \leq j \leq n.$$

Summing over j we get

2010 Mathematics Subject Classification: 15A39, 97H30
Key words and phrases: Cauchy-Schwarz inequality, Heisenberg uncertainty principle, Cramer-Rao inequality.

$$\sum_{j=1}^n 2a_jb_j \leq \lambda^2 \sum_{j=1}^n a_j^2 + \lambda^{-2}\sum_{j=1}^n b_j^2.$$

Choosing

$$\lambda^2 = \left(\sum_{j=1}^n b_j^2\right)^{1/2}\left(\sum_{j=1}^n a_j^2\right)^{-1/2}$$

so as to minimize the right hand side leads immediately to the Cauchy-Schwarz inequality.

The result is easily extended to complex numbers a_i, b_i by observing that

$$\left|\sum_{j=1}^n a_jb_j\right| \leq \sum_{j=1}^n |a_j||b_j| \tag{1.4}$$

and then applying (1.1) to the latter sum so as to deduce

$$\left|\sum_{j=1}^n a_jb_j\right| \leq \left(\sum_{j=1}^n |a_j|^2\right)^{1/2}\left(\sum_{j=1}^n |b_j|^2\right)^{1/2}.$$

It is not difficult to see that again, equality can arise if and only if (1.3) holds, because (1.4) is a consequence of the triangle inequality in which the case of equality is easily identified.

The corresponding inequality for integrals, namely

$$\left|\int_a^b f(x)g(x)dx\right| \leq \left(\int_a^b |f(x)|^2dx\right)^{1/2}\left(\int_a^b |g(x)|^2dx\right)^{1/2} \tag{1.5}$$

seems to have been first stated by Buniakowsky in 1859 and later (independently) by Schwarz in 1885 and for this reason, we sometimes refer to this as the Cauchy-Buniakowsky-Schwarz inequality (see page 16 of [6]). It is easy to deduce (1.5) from (1.1) by using Riemann sums and the limiting process. It is also easy to see that (1.5) extends to improper integrals. We leave the details to the student.

All these inequalities are special cases of a more general theorem:

Theorem 1. *If V is an inner product space over $\mathbb{R}$ or $\mathbb{C}$, then*

$$|(v,w)| \leq ||v||||w||$$

for all $v, w \in V$ with equality if and only if v is a scalar multiple of w.

Proof. We may suppose that $w \neq 0$, for otherwise, the result is clear. We decompose v into its components parallel and perpendicular to w by writing: $v = \lambda w + (v - \lambda w)$ for some scalar λ. For $v - \lambda w$ to be perpendicular to w, we need $0 = (v - \lambda w, w) = (v,w) - \lambda||w||^2$, that is, $\lambda = (v,w)/||w||^2$. Now, $(v - \lambda w, v - \lambda w) \geq 0$ for any λ so that $||v||^2 - \overline{\lambda}(v,w) - \lambda(w,v) + |\lambda|^2||w||^2 \geq 0$. With our choice of λ in particular, we deduce

$$||v||^2 - 2\frac{|(v,w)|^2}{||w||^2} + \frac{|(v,w)|^2}{||w||^4}||w||^2 \geq 0.$$

In other words, $|(v,w)| \leq ||v||||w||$. Clearly, equality can occur if and only if $v = \lambda w$ for some λ. This completes the proof. □

Inequalities (1.1) and (1.5) are now special cases of this more general inequality using the appropriate inner product spaces such as $L^2[a,b]$.

2. A principle of duality

At the center of sieve theory and the large sieve inequality in particular, lies a fundamental principle of duality which is essentially the Cauchy-Schwarz inequality. We record this below.

Theorem 2. *Let c_{ij}, for $1 \leq i \leq m$, $1 \leq j \leq n$, be mn complex numbers. Let λ be a non-negative real number. Then, the inequality*

$$\sum_{i=1}^{m} \left| \sum_{j=1}^{n} c_{ij} a_j \right|^2 \leq \lambda \sum_{j=1}^{n} |a_j|^2$$

holds for all complex numbers $a_1, ..., a_n$ if and only if the inequality

$$\sum_{j=1}^{n} \left| \sum_{i=1}^{m} c_{ij} b_i \right|^2 \leq \lambda \sum_{i=1}^{m} |b_i|^2$$

holds for all complex numbers $b_1, ..., b_m$.

Proof. Let C be the $m \times n$ matrix (c_{ij}) and $\mathbf{a} = (a_1, ..., a_n)^{\mathrm{tr}}$ and $\mathbf{b} = (b_1, ..., b_m)^{\mathrm{tr}}$ be column vectors in $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively. The first inequality of the theorem can then be written as $(C\mathbf{a}, C\mathbf{a}) \leq \lambda(\mathbf{a}, \mathbf{a})$, and the second one as $(\mathbf{b}^{\mathrm{tr}}C, \mathbf{b}^{\mathrm{tr}}C) \leq \lambda(\mathbf{b}, \mathbf{b})$. Suppose the first inequality holds and let $\mathbf{b} \in \mathbb{C}^m$. Then, by the Cauchy-Schwarz inequality, we have for all $\mathbf{a} \in \mathbb{C}^n$,

$$(\mathbf{b}^{\mathrm{tr}}C\mathbf{a}, \mathbf{b}^{\mathrm{tr}}C\mathbf{a}) \leq ||\mathbf{b}||^2||C\mathbf{a}||^2 \leq \lambda||\mathbf{b}||^2||\mathbf{a}||^2$$

by our assumption. Now set $\mathbf{a}{=}\overline{C}^{\mathrm{tr}}\mathbf{b}$ to deduce $||\mathbf{b}^{\mathrm{tr}}C||^4 \leq \lambda||\mathbf{b}||^2||\mathbf{b}^{\mathrm{tr}}C||^2$ which gives the result. The converse is similarly deduced. □

The principle of duality can be used to deduce what is called the large sieve inequality which plays a fundamental role in analytic number theory. The reader can find further details in the monograph [1] as well as [3].

3. The Heisenberg uncertainty principle

The celebrated Heisenberg uncertainty principle, which is a corner stone of quantum mechanics, is an immediate consequence of the Cauchy-Schwarz inequality once we understand the dictionary that translates the concepts of physics into mathematical language. Indeed, from a mathematical standpoint, Heisenberg's uncertainty principle states that a function F and its Fourier transform $\widehat{F}$ cannot **both** have compact support. To prove this, it is convenient to introduce the following class of functions.

We define the Schwartz space (after Laurent Schwartz and no relation to the Schwarz of the Cauchy-Schwarz inequality) $\mathcal{S}$ to be the space of infinitely differentiable functions $F : \mathbb{R} \to \mathbb{C}$ such that $|x^k F^{(\ell)}(x)| \to 0$ as $|x| \to \infty$, for all non-negative integers k and ℓ.

We recall the definition of the Fourier transform $\widehat{\psi}$ of a function $\psi \in \mathcal{S}$:

$$\widehat{\psi}(x) := \int_{-\infty}^{\infty} \psi(t) e^{-2\pi i t x} dt, \quad x \in \mathbb{R}.$$

It is easily verified that $\widehat{\psi}$ is also in $\mathcal{S}$.

Theorem 3 (Heisenberg's uncertainty principle)**.** *Let ψ be a Schwartz function satisfying $||\psi||_2 = 1$. Then,*

$$\left(\int_{-\infty}^{\infty} x^2 |\psi(x)|^2 dx\right)\left(\int_{-\infty}^{\infty} x^2 |\widehat{\psi}(x)|^2 dx\right) \geq \frac{1}{16\pi^2},$$

with equality if and only if $\psi(x) = Ae^{-Bx^2}$ for some $B > 0$ and $|A|^2 = \sqrt{\frac{2B}{\pi}}$.

Proof. Writing

$$1 = \int_{-\infty}^{\infty} |\psi(x)|^2 dx = \int_{-\infty}^{\infty} \psi(x)\overline{\psi(x)} dx,$$

we integrate by parts to get that this is

$$= x\psi(x)\overline{\psi(x)}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} x \frac{d}{dx}\left(\psi(x)\overline{\psi(x)}\right) dx.$$

Because $\psi \in \mathcal{S}$, the first term equals zero giving us

$1 = -\int_{-\infty}^{\infty} x\{\psi(x)\overline{\psi'(x)} + \psi'(x)\overline{\psi(x)}\} dx$, and hence $1 \leq 2\int_{-\infty}^{\infty} |x\psi(x)\psi'(x)| dx$.

Applying the Cauchy-Schwarz inequality, we get

$$1 \leq 2\left(\int_{-\infty}^{\infty} x^2 |\psi(x)|^2 dx\right)^{1/2}\left(\int_{-\infty}^{\infty} |\psi'(x)|^2 dx\right)^{1/2}.$$

By the Fourier inversion theorem,

$$\psi(x) = \int_{-\infty}^{\infty} \widehat{\psi}(t) e^{2\pi i t x} dt, \text{ so that } \psi'(x) = \int_{-\infty}^{\infty} (2\pi i t)\widehat{\psi}(t) e^{2\pi i t x} dt,$$

the differentiation under the integral sign being justified by the virtues of the elements of the Schwartz class $\mathcal{S}$. In other words, $\psi'(-x)$ is the Fourier transform of $(2\pi i t)\widehat{\psi}(t)$. By Parseval's formula, we deduce that the L^2-norm of ψ' is equal to $\int_{-\infty}^{\infty} 4\pi^2 t^2 |\widehat{\psi}(t)|^2 dt$. Thus,

$$1 \leq 4\pi\left(\int_{-\infty}^{\infty} x^2 |\psi(x)|^2 dx\right)^{1/2}\left(\int_{-\infty}^{\infty} x^2 |\widehat{\psi}(x)|^2 dx\right)^{1/2},$$

from which the main inequality emerges. For the final part of the theorem, we note that in our application of the Cauchy-Schwarz inequality, equality

can occur if and only if $\psi'(x) = \lambda x\psi(x)$, for some scalar λ. This is an ordinary differential equation which is easily solved. We find $\psi(x) = Ae^{-Bx^2}$ for certain constants A and $B > 0$ because ψ belongs to $\mathcal{S}$. The fact that the L^2-norm of ψ equals 1 gives us the final claim. □

An immediate consequence of the uncertainty principle is that ψ and $\widehat{\psi}$ cannot both be concentrated in a small neighborhood of the origin. Indeed, suppose ψ is supported in $[-M, M]$ and $\widehat{\psi}$ is supported in $[-N, N]$. Then, $M^2N^2 \geq 1/4\pi^2$. So, M and N cannot both be arbitrarily small. A manifestation of the uncertainty principle is what we mentioned earlier that both ψ and $\widehat{\psi}$ cannot have compact support. If ψ has compact support, then its Fourier transform is an entire function (see for example, pp. 371-372 of [11]). As such, the zeros of $\widehat{\psi}$ are isolated unless it is identically zero, in which case ψ is also identically zero by the inversion theorem.

In quantum mechanics, $|\psi(x)|^2dx$ represents the probability density that a particle is near position $x \in \mathbb{R}$. Thus, the probability that such a particle lies in the interval $[a, b]$ is given by

$$\int_a^b |\psi(x)|^2dx.$$

Our best guess for the position of the particle is given by the expectation

$$\mu := \int_{-\infty}^{\infty} x|\psi(x)|^2dx$$

and the error (or uncertainty) involved in this guess is given by the variance

$$\int_{-\infty}^{\infty} (x-\mu)^2|\psi(x)|^2dx.$$

A similar analysis holds for the momentum of the particle. Indeed, the probability density that the momentum is $x \in \mathbb{R}$ is $|\widehat{\psi}(x)|^2dx$ and so the probability that the momentum lies in $[a, b]$ is given by

$$\int_a^b |\widehat{\psi}(x)|^2dx.$$

The expectation and variance of the momentum are defined as before. Without any loss of generality, we can normalize our functions so that the expectation of both the position and momentum are zero. With this normalization, we see that Heisenberg's uncertainty principle establishes a lower bound for the product of these variances (or errors). In other words, any attempt to lower the error in our observation of the position of a particle increases the error in determining its momentum and vice versa.

An uncertainty principle of some sort or other abounds in nature described by Fourier duality. Recently, Tao [13] discovered a discrete version of the uncertainty principle and a simple proof of this can be found in [8]. See also [7] for other variations on this theme. An uncertainty principle for the equidistribution of arithmetic sequences in arithmetic progressions and short intervals has been a topic of intense research in analytic number theory. The reader can find an exposition of this in [5].

4. The Cramér-Rao inequality

A humorous definition of statistics is that it is the converse of probability theory. Though seemingly funny, the joke contains the essential idea. In probability theory, we deal with measurable functions (also called random variables), probabilty density functions (sometimes referred to as pdfs or probability measures). By contrast, in statistics, we may not know what the pdf may be of a certain phenomenon. At best, we can take a large sample and infer from this data, the nascent pdf that describes the phenomenon.

At the dawn of the 20th century, R.A. Fisher undertook the task of laying the foundations of theoretical statistics. In his fundamental paper [4], he wrote "the object of statistical methods is the reduction of data." Expanding on this aphorism, he explained, "A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole." Motivated by these considerations, he was led to define (what we now call) the *Fisher information* of a random variable X with probability density function $f_\theta(x)$ attached to an unknown deterministic parameter θ as follows. The contribution of x to the information content of the random variable may be viewed to be $-\log(f_\theta(x))$ since from information theory we know that for the discrete setup an ideal lossless binary code would need roughly these many bits to represent this variable.

The rate of change of this information is then $\frac{\partial \log f_\theta}{\partial \theta}$ and it seems reasonable that the expectation

$$I(\theta) := \int_{-\infty}^{\infty} \left(\frac{\partial \log f_\theta}{\partial \theta}\right)^2 f_\theta(x)dx$$

(called *Fisher information* in statistical parlance) gives us some idea of the amount of information about θ contained in the data.

For example, if X has a Bernoulli distribution where X can have only two values "heads" or "tails" (or more precisely 0 and 1 say), with 1 having

probability θ and 0 having probability $1-\theta$, with density function

$$f_\theta(x) = \theta^x(1-\theta)^{1-x},$$

then $\log f_\theta(x) = x\log\theta + (1-x)\log(1-\theta)$ and we find, after a simple calculation, that $I(\theta) = (\ \theta(1-\theta)\ )^{-1}$.

To take another example, suppose X has a normal distribution with unknown mean μ and known variance σ^2. Let us determine the Fisher information $I(\mu)$ in X. Since $f_\mu(x) = (\sqrt{2\pi}\,\sigma)^{-1}\, e^{-(x-\mu)^2/2\sigma^2}$, we see $\log f_\mu(x) = -(1/2)\log(2\pi\sigma^2) - (x-\mu)^2/2\sigma^2$, so that $\dfrac{\partial \log f_\mu}{\partial \mu} = (x-\mu)/\sigma^2$. A direct calculation now gives that $I(\mu) = 1/\sigma^2$.

In statistical problems, large amounts of data are collected to study a phenomenon. With a desire to derive a mathematical model to describe it, we may find, numerically, a function $\widetilde{\phi}$ to approximate a parameter ϕ. $\widetilde{\phi}$ is called an *unbiased estimator* of ϕ if $E(\widetilde{\phi}) = \phi$. That is,

$$\int_{-\infty}^{\infty} \widetilde{\phi} f_\theta(x)dx = \phi(\theta).$$

Here, θ and x are independent parameters. Differentiating this with respect to θ and interchanging integration and differentiation (provided of course that this is permissible) gives:

$$\int_{-\infty}^{\infty} \widetilde{\phi}(x)\frac{\partial f_\theta}{\partial \theta}(x)dx = \phi'(\theta).$$

The rate of change of information is the function

$$S(x) := \frac{\partial}{\partial \theta}\log f_\theta(x)$$

called the *score statistic*. Plainly, $S(x) = \frac{1}{f_\theta(x)}\frac{\partial f_\theta}{\partial \theta}(x)$, so that we can write

$$\int_{-\infty}^{\infty} \widetilde{\phi}(x)S(x)f_\theta(x)dx = \phi'(\theta). \tag{4.1}$$

Also, the expectation of $S(x)$ is

$$E(S(x)) = \int_{-\infty}^{\infty} S(x)f_\theta(x)dx = \int_{-\infty}^{\infty} \frac{\partial f_\theta}{\partial \theta}(x)dx = \frac{\partial}{\partial \theta}\int_{-\infty}^{\infty} f_\theta(x)dx = 0,$$

since

$$\int_{-\infty}^{\infty} f_\theta(x)dx = 1,$$

because the total probability is 1. Thus, (4.1) can be re-written as

$$\int_{-\infty}^{\infty} (\widetilde{\phi}(x) - \phi(\theta))S(x)f_\theta(x)dx = \phi'(\theta).$$

Applying the Cauchy-Schwarz inequality, we obtain

$$\phi'(\theta)^2 \le \left(\int_{-\infty}^{\infty} (\widetilde{\phi}(x) - \phi(\theta))^2 f_\theta(x)dx\right)\left(\int_{-\infty}^{\infty} S(x)^2 f_\theta(x)dx\right).$$

Writing

$$I(\theta) := \int_{-\infty}^{\infty} \left(\frac{\partial \log f_\theta}{\partial \theta} \right)^2 f_\theta(x) dx,$$

(called *Fisher information* in statistical parlance), we can write our inequality as:

Theorem 4 (The Cramér-Rao inequality)**.** *For an unbiased estimator $\widetilde{\phi}$ of ϕ, we have*

$$\int_{-\infty}^{\infty} (\widetilde{\phi}(x) - \phi(\theta))^2 f_\theta(x) dx \geq \frac{\phi'(\theta)^2}{I(\theta)}.$$

Often, this is applied with $\phi(\theta) = \theta$ so that $\phi'(\theta) = 1$. The inequality then gives us a limitation on the accuracy of the unbiased estimator to the function θ. Somtimes it is referred to as the information inequality. It was discovered independently by C. R. Rao [10] and H. Cramér [2] in 1945 and has played a pivotal role in statistical inference. An enlightening survey of the Cramér-Rao inequality was written by K.R. Parthasarathy [9] where the reader can find discussion of Riemannian metrics to study population models.

Regarding Theorem 4, there is a lot of interest in estimators that actually achieve the Cramer-Rao lower bound. Such estimators are said to be asymptotically efficient. Under certain regularity conditions the maximum likelihood estimators are asymptotically efficient. In such cases the Fisher information about θ in the data is equal to the inverse of the variance of the estimator.

Acknowledgements. I would like to thank Devon Lin, Cyrus Mehta, Neha Prabhu, Siddhi Pathak, François Séguin, Serdar Yuksel and the anonymous referee for their comments on an earlier version of this paper.

References

[1] Cojocaru, A. and Ram Murty, M., *An introduction to sieve methods and their applications*, London Mathematical Society Student Texts 66, Cambridge University Press, 2006.

[2] Cramér, H., *Mathematical methods of statistics*, Princeton University Press, Princeton, 1946.

[3] Elliott, P. T. D. A., On inequalities of large sieve type,*Acta Arith.*, **18** (1971), 405–422.

[4] Fisher, R. A., On the mathematical foundations of theoretical statistics, *Phil. Trans. Royal Soc. London*, Ser. A, **222A** (1922), 309–368.

[5] Granville, A. and Soundararajan, K., An uncertainty principle for arithmetic sequences, *Annals of Math.*, **165** (2007), 593–635.

[6] G.H. Hardy, G. H., Littlewood, J. E. and Pólya, G., *Inequalities*, Cambridge University Press, 1967.

[7] Ram Murty, M., Some remarks on the discrete uncertainty principle, in Highly Composite, Papers in Number Theory, *RMS Lecture Notes Series*, **23** (2016), 77–85.

[8] Ram Murty, M. and Whang, P., The uncertainty principle and a generalization of a theorem of Tao, *Linear Algebra and Applications*, **437** (2012), no. 1, 214–220.

[9] Parthasarathy, K. R., On the philosophy of Cramér-Rao-Bhattacharya inequalities in quantum statistics, in *Frontiers of Inference in Statistics and Sciences*, Vol. **1** (2010), 198–220; edited by S. B. Rao, C. R. Rao, Advanced Institute for Mathematics, Statistics and Computer Science, Hyderabad.

[10] Rao, C. R., Information and the accuracy attainable in the estimation of statistical parameters, *Bull. Calcutta Math. Soc.*, **37** (1945), 81–91.

[11] Rudin, W., *Real and Complex Analysis*, Third edition, McGraw-Hill, Boston, 1987.

[12] Stein, E. and Shakarchi, R., *Fourier analysis, an introduction*, Princeton University Press, Princeton, 2003.

[13] Tao, T., An uncertainty principle for cyclic groups of prime order, *Math. Res. Lett.*, **12** (1) (2005), 121–127.

M. Ram Murty
Dept. of Mathematics and Statistics
Queen's University, Kingston, Ontario, Canada, K7L 3N6
E-mail: murty@queensu.ca