Inequalities in IA Probability

Richard Chapling v1.1 18 March 2020

In this handout we consider the two types of inequality for random variables of interest in this course: tail bounds and expectation bounds. (We will not we discussing the basic inequalities that $\mathbb{P}$ satisfies such as subadditivity and the Bonferroni inequalities.)

1 Tail bounds

Many useful probability distributions have formulae that make calculating their various properties exactly rather difficult: even if we know some formulae, other properties may not be expressible in a way that is simple enough to give us useful information. You have already seen this in the case of the binomial distribution: the distribution function does not have a closed form. However, it is often possible to obtain some information about a distribution's properties by *approximating* them.

In this section, we are interested in tail bounds: these are used to establish upper estimates on the probability that X deviates from some point, usually its mean. They are useful in proving results such as the Law of Large Numbers and the Central Limit Theorem.

The examples we look at here take the "global" information contained in the expectation of some $f(X)$, which is normally reasonably easy to calculate, and give estimates of the probability that X is larger than some given value. Because these bounds must be true for any functions where the appropriate expectations are finite, they are inevitably somewhat crude: the art often lies in finding a sufficiently rapidly-growing function f that that the expectation of $f(X)$ being finite is genuinely an effective condition.

On the other hand, *because* these bounds are weak, anything that we can prove using them tells us something about a lot of random variables all at once. The Law of Large Numbers is a good example of this: the conditions in the theorem are really quite weak, so it applies to a lot of random variables.

1.1 Markov and Chebyshev's inequalities

The exemplar of this sort of bound is

Theorem 1 (Markov's inequality[1]). *Let X be a random variable with $X \geqslant 0$ and $\mathbb{E}[X] < \infty$. Then for any $a > 0$,*

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

(In fact if $\mathbb{E}[X] = \infty$, the inequality is still true, but is useless.)

Proof. Since $X \geqslant 0$, we have

$$a\mathbb{1}\{X \geqslant a\} \leqslant X\mathbb{1}\{X \geqslant a\} \leqslant X(\mathbb{1}\{X \geqslant a\} + \mathbb{1}\{X < a\}) = X,$$

[1] The original bound of this sort is due to Chebyshev, see below. Markov published something more like this form rather later [19, p. 65 of the original], [18, p. 54 of the 1912 German translation]. The reader will rapidly learn that most inequalities, like most theorems, have the wrong names attached to them.

and applying expectations to both sides,

$$a\mathbb{E}[\mathbb{1}\{X \geqslant a\}] \leqslant \mathbb{E}[X].$$

Since $\mathbb{E}[\mathbb{1}\{A\}] = \mathbb{P}(A)$ and $\mathbb{E}[X] < \infty$, the result follows. □

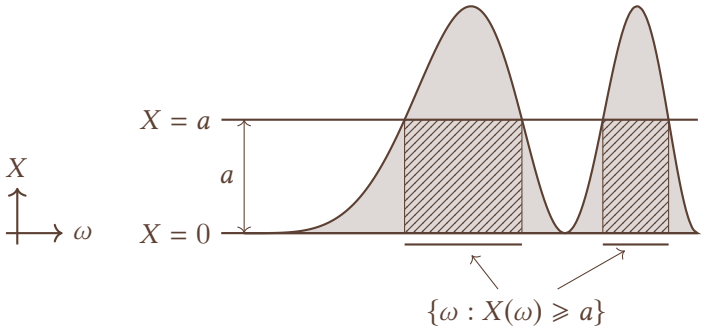


Figure 1: The picture to remember Markov's inequality: the shaded part has area $\mathbb{E}[X]$, the hatched part area $a\mathbb{P}(X \geqslant a)$. (Note in this example Ω and X has been set up to look very simple: this diagram is schematic and mnemonic rather than realistic.)

Markov's inequality is *the weakest possible bound* if $\mathbb{E}[X]$ is finite. Its power lies in applying it to well-chosen random variables with finite expectation, rather than applying it to X itself. A good example of this is

Corollary 2 (Chebyshev's inequality[2]). *Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then for any $a > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) \leqslant \frac{\operatorname{Var} X}{a^2}$$

Proof. The left-hand side is $\mathbb{P}((X - \mathbb{E}[X])^2 \geqslant a^2)$. But since $\mathbb{E}[X^2] < \infty$, $(X - \mathbb{E}[X])^2$ is a nonnegative random variable with finite expectation, so Markov applies and the result falls out. □

1.2 Further consequences of Markov's inequality

To finish this section, we give some more examples of extensions of Markov's inequality to illustrate our earlier comment about skilful application.

Corollary 3 (General tail estimate). *Let f be a nondecreasing nonnegative function. If X is any random variable with $\mathbb{E}[f(X)] < \infty$,[3] then for any $a > 0$,*

$$\mathbb{P}(X \geqslant a) = \mathbb{P}(f(X) \geqslant f(a)) \leqslant \frac{\mathbb{E}[f(X)]}{f(a)}.$$

Proof. As in the original proof of Markov's inequality, we have

$$f(a)\mathbb{1}\{X \geqslant a\} \overset{f \text{ incr}}{=} f(a)\mathbb{1}\{f(X) \geqslant f(a)\} \overset{f \text{ incr}}{\leqslant} f(X)\mathbb{1}\{f(X) \geqslant f(a)\} \overset{f(X)\geqslant 0}{\leqslant} f(X),$$

and taking expectations gives the result. □

[2] Originally proven by Chebyshev [28], but originally stated in 1853 by Bienaymé (see the paper before Chebyshev's, which is a reprint of Bienaymé's paper).

[3] In particular X is now allowed to be negative with nonzero probability.

Corollary 4 (Moment bound). *Let $p \geqslant 0$. If X is any random variable with $\mathbb{E}[|X|^p] < \infty$, for any $a > 0$,*

$$\mathbb{P}(|X| \geqslant a) \leqslant \mathbb{E}[X^p]a^{-p}.$$

Corollary 5 (MGF bound). *Let $\beta \geqslant 0$. If X is any random variable with $\mathbb{E}[e^{tx}] < \infty$, for any $a > 0$,*

$$\mathbb{P}(X \geqslant a) \leqslant \mathbb{E}[e^{tX}]e^{-ta}.$$

Existence of the MGF $\mathbb{E}[e^{tx}]$ is a very strong condition, because it requires exponential decay in the tail. Most useful probability distributions you meet after this course do not do this!

By minimising the right-hand side over t, we obtain a simple bound that tends to be of great strength, known as a *Chernoff bound*.[4] While not as strong as the best moment bound, it is normally much easier to compute.[5] For example, we know that if $X \sim \mathrm{N}(\mu, \sigma^2)$, $m(\theta) = \mathbb{E}[e^{\theta X}] = \exp(\mu\theta + \sigma^2\theta^2/2)$, and we compute that if $a \geqslant \mu$,

$$\mathbb{P}(X \geqslant a) \leqslant \inf_{\theta \geqslant 0} m(\theta)e^{-\theta x} = \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right),$$

which is close to the bound $\frac{1}{\sqrt{2\pi}t}\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right)$ which we can obtain by scrutinising the integral more closely. On the other hand, computing the moment bound requires the Gamma-function, and minimising this would require being able to invert the derivative of the Digamma function.

Lastly, we consider when we have equality in Markov's inequality. We need the following lemma:

Lemma 6. *If $X \geqslant 0$, then $\mathbb{E}[X] = 0 \iff \mathbb{P}(X = 0) = 1$.*

Proof. $\Longleftarrow$ is clear. For $\Longrightarrow$, for any $a > 0$, by Markov's inequality,

$$0 = \mathbb{E}[X] \geqslant a\mathbb{P}(X \geqslant a) \geqslant 0,$$

so $\mathbb{P}(X \geqslant a) = 0$. But $\{X > 0\} = \bigcup_{n=1}^{\infty}\{X \geqslant 1/n\}$, so by Boole's inequality,

$$\mathbb{P}(X > 0) \leqslant \sum_{n=1}^{\infty}\mathbb{P}(X \geqslant 1/n) = 0,$$

as required. □

Corollary 7. *Equality occurs in Markov's inequality if and only if $\mathbb{P}(X \in \{0, a\}) = 1$.*

Proof. Let $Y = X - a\mathbb{1}\{X \geqslant a\}$. We showed in the proof of Markov that Y is nonnegative, and we see that $\mathbb{E}[Y]$ is the difference between the sides of Markov's inequality, so it is nonnegative. Hence by the lemma $\mathbb{P}(Y = 0) = 1$, so $X = a\mathbb{1}\{X \geqslant a\}$ with probability 1. There are then two cases: either $\mathbb{1}\{X \geqslant a\} = 0$, in which case $X = 0$, or $\mathbb{1}\{X \geqslant a\} = 1$, in which case $X = a$. So $\mathbb{P}(X \in \{0, a\}) = 1$. □

[4] This appears in Chernoff's work [8], but Chernoff attributes the idea to Herman Rubin [1]:

> Rubin claimed that part of my derivation, giving the lower bounds, could be obtained much more easily. After working so hard, I doubted it very much. He showed me the Chebyshev type of proof that gives rise to what's now called the Chernoff bound, but it is certainly Rubin's. When I wrote up the technical report, I mentioned his assistance but when I submitted the paper for publication, I left it out because it was so trivial and it never occurred to me that this would be one of the things that would lead to my fame in electrical engineering circles. That inequality turned out to be a very important result as far as information theory is concerned, and so the lower bound has been called the Chernoff bound ever since. I am very unhappy about the fact that I did not properly credit Rubin at that time because I thought it was a rather trivial lemma, but many things are only trivial once you know them.

There is a moral about proper citation here.

[5] See, e.g. [21] for a thorough survey.

Rather pleasantly, we were able to use Markov itself to obtain the condition for its own equality case!

For example, let $X \sim \mathrm{Ber}(p)$, and let $0 < a \leqslant 1$. Then $\mathbb{P}(X \geqslant a) = p = \mathbb{E}[X]$, so $p \leqslant p/a$ and we obtain equality if and only if $a = 1$, as expected (if $a > 1$, $\mathbb{P}(X \geqslant a) = 0$ and we never get equality).

We shall not proceed further with this: the reader interested in extensions of the very simple beginning of the theory outlined here is directed to the extensive statistical literature on tail bounds and concentration inequalities.

2 Convex functions and Jensen's inequality

In this section we prove a general inequality about functions, which can be used to derive both many classical inequalities, and counterparts that are useful for bounding expectations of random variables in terms of expectations of others.

2.1 Convex functions

In the following, I is always an interval $\subseteq \mathbb{R}$.

Definition 8. A function $\varphi\colon I \to \mathbb{R}$ is called *convex*[6] if

$$(\forall x, y \in I)(\forall \lambda \in (0,1)) \qquad (1-\lambda)\varphi(x) + \lambda\varphi(y) \geqslant \varphi\big((1-\lambda)x + \lambda y\big).$$

If the inequality is strict $>$, φ is called *strictly convex*.

That is, φ is convex if no chord goes below the graph of φ, and strictly convex if the interior of the chord is always above the graph of φ.

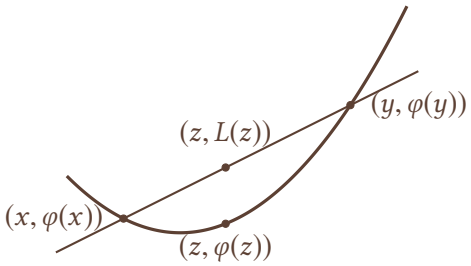


Figure 2: The picture to remember for convexity: the point $(z, L(z))$ lies above the point $(z, \varphi(z))$, where L is the line through $(x, \varphi(x))$ and $(y, \varphi(y))$ and z is between x and y.

For this course, we need to know two things about convex functions:

Proposition 9 (Second-order sufficient condition for convexity). *Let $\varphi\colon I \to \mathbb{R}$ be twice-differentiable. If $\varphi'' \geqslant 0$, φ is convex. If $\varphi'' > 0$, φ is strictly convex.*

In fact this condition is also necessary, but we are not interested in characterising convex functions: we just want an easy way to prove that certain functions are convex so that we can use them.

Proof. Let $x, y, z \in I$ with $x < z < y$. Since φ is differentiable, it is continuous on $[x, z]$ and differentiable on (x, z). By the Mean Value Theorem applied to φ', $\exists \zeta \in (x, z)$ so that

$$\varphi(x) - \varphi(z) = \varphi'(\zeta)(x - z),$$

and similarly, $\exists \eta \in (z, y)$ so that

$$\varphi(y) - \varphi(z) = \varphi'(\eta)(y - z),$$

[6] This term was first introduced by Jensen [13, p. 176], in the same paper as his inequality discussed below.

Adding $(1-\lambda)$ of the first to λ of the second,

$$\big((1-\lambda)\varphi(x)+\lambda\varphi(y)\big)-\varphi(z)=(1-\lambda)\varphi'(\zeta)(x-z)+\lambda\varphi'(\eta)(y-z).$$

Putting $z=(1-\lambda)x+\lambda y$, where $\lambda\in(0,1)$, we have $x-z=-\lambda(y-x)$ and $y-z=(1-\lambda)(y-x)$, so the right-hand side becomes

$$\lambda(1-\lambda)(y-x)(\varphi'(\eta)-\varphi'(\zeta))$$

Finally, since φ is twice-differentiable, φ' exists and is continuous on $[\zeta,\eta]$ and differentiable on (ζ,η). By the Mean Value Theorem applied to φ', $\exists\xi\in(\zeta,\eta)$ so that

$$\varphi'(\eta)-\varphi'(\zeta)=\varphi''(\xi)(\eta-\zeta),$$

so

$$\big((1-\lambda)\varphi(x)+\lambda\varphi(y)\big)-\varphi((1-\lambda)x+\lambda y)=\lambda(1-\lambda)(y-x)\varphi''(\xi)(\eta-\zeta)\geqslant 0$$

which is $\geqslant 0$ if $\varphi''\geqslant 0$, or >0 if $\varphi''>0$. □

The reader will notice that the proof also implies that φ is convex if φ' is increasing. This is a more general condition, but slightly harder to check.

The other property of convex functions that we will want is that we can find a linear function lying under them. When the function is differentiable, the tangent will do:

Proposition 10. *Suppose that $\varphi\colon I\to\mathbb{R}$ is convex and differentiable. Then for any $z\in I$,*

$$(\forall x\in I)\qquad \varphi(x)\geqslant\varphi(z)+\varphi'(z)(x-z);$$

that is, φ does not go below its tangent at z. If φ is strictly convex, the inequality is strict for all $x\in I\setminus\{z\}$.

Proof. The easiest way to do this is to use that φ' is nondecreasing, along with the MVT. By the MVT, there is ξ between z and x so that

$$\varphi(x)-\varphi(z)=\varphi'(\xi)(x-z)=(\varphi'(\xi)-\varphi'(z))(x-z)+\varphi'(z)(x-z).$$

But φ' is nondecreasing, so since the sign of $x-z$ is the same as the sign of $\xi-z$, and hence $\varphi'(\xi)-\varphi'(z)$ is either 0 or has the same sign as $x-z$, so the first term is always nonnegative and the result follows. If φ is strictly convex, φ' is strictly increasing, so the first term is always positive if $x\neq z$. □

Support lines Most of the time, your convex functions will be differentiable. For situations where they are not, here we give a brief discussion of the more general situation. We introduce a new notion:

Definition 11 (Support line). Let $\varphi\colon I\to\mathbb{R}$. A linear function L is called a *support line* for φ if for each $x\in I$ we have $\varphi(x)\geqslant L(x)$, with at least one point $z\in I$ where equality is achieved.

It is apparent that this can act as a replacement for the use of the tangent line as a linear function that lies below a convex function. First we need to prove that these things actually exist.

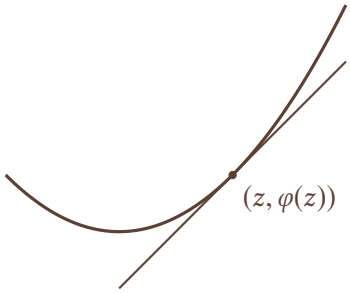
Proposition 12 (Convex functions have support lines). *Let $\varphi\colon I\to\mathbb{R}$ be convex and $z\in I$. Then there is a support line for φ that passes through $(z,\varphi(z))$.*

Proof. Let $x<z<y$. Then the convexity condition says

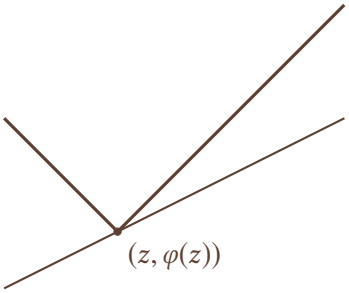
$$\varphi((1-\lambda)x+\lambda y)\leqslant(1-\lambda)\varphi(x)+\lambda\varphi(y),$$

or

$$\frac{\varphi((1-\lambda)x+\lambda y)-\varphi(x)}{\lambda}\leqslant\frac{\varphi(y)-\varphi((1-\lambda)x+\lambda y)}{1-\lambda}.$$



(a) Differentiable convex function and supporting tangent line



(b) Support line at a point where φ is not differentiable

Figure 3: Convex functions φ with support lines

Taking $\lambda = (z - x)/(y - x)$ gives

$$\frac{\varphi(z) - \varphi(x)}{z - x} \leqslant \frac{\varphi(y) - \varphi(z)}{y - z}.$$

Since this is true for any $x < z < y$,

$$\sup_{x<z} \frac{\varphi(z) - \varphi(x)}{z - x} \leqslant \inf_{y>z} \frac{\varphi(y) - \varphi(z)}{y - z},$$

so there is at least one real number m between these two. Write $L(w) := m(w - z) + \varphi(z)$. Then $L(z) = \varphi(z)$ and

$$\varphi(w) - L(w) = (\varphi(w) - \varphi(z)) - m(w - z).$$

If $w < z$, $m \geqslant \sup_{x<z} \frac{\varphi(z)-\varphi(x)}{z-x}$, so $-m(w - z) \geqslant \varphi(z) - \varphi(w)$ and so $\varphi(w) - L(w) \geqslant 0$, and similarly if $w > z$. Hence L is the required support line. □

Lastly, we need to worry about equality cases. The easiest way around this is to demand strict convexity:

Lemma 13. *If φ is strictly convex, no support line intersects φ at more than one point.*

Proof. Suppose that φ is convex and L is a support line for φ, intersecting φ at distinct points $x < y$. Then $g = \varphi - L$ is a sum of two convex functions and so convex. We have $g(z) = \varphi(z) - L(z) \geqslant 0$ for any $x < z < y$. But $g(x) = g(y) = 0$, so by convexity,

$$g(z) \leqslant \frac{y - z}{y - x} g(x) + \frac{z - x}{y - x} g(y) = 0.$$

Hence $g(z) = 0$, and it follows, since $L(z) = \frac{y-z}{y-x} L(x) + \frac{z-x}{y-x} L(y)$, that $\varphi(z) = \frac{y-z}{y-x} \varphi(x) + \frac{z-x}{y-x} \varphi(y)$, so φ cannot be strictly convex. □

Therefore if φ is strictly convex and L is a support line for φ through $(z, \varphi(z))$, $\varphi(x) > L(x)$ for every $x \neq z$.

The main result in this section is the following:

Theorem 14 (Jensen's inequality[7]). *Let X be a real-valued random variable, and φ a (differentiable) convex function. Then*

$$\mathbb{E}[\varphi(X)] \geqslant \varphi(\mathbb{E}[X]).$$

If φ is strictly convex on an interval containing $\mathbb{E}[X]$, equality occurs if and only if $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

This is not the most general condition for equality, but again will be sufficient for what we do in this course.

[7] Again, while this inequality is universally named after Jensen [13], it had been proven years earlier by Hölder [12] for φ with $\varphi'' > 0$.

Proof. Putting $z = \mathbb{E}[X]$ in either Proposition 10 or Proposition 12,

$$\varphi(X) \geqslant \varphi(\mathbb{E}[X]) + m(X - \mathbb{E}[X]), \tag{1}$$

and the result follows since expectation preserves nonstrict inequalities. If φ is strictly convex at $\mathbb{E}[X]$, either the last part of Proposition 10, or more generally Lemma 13 imply that the inequality (1) is strict for every $X \neq \mathbb{E}[X]$, so if $\mathbb{P}(X \neq \mathbb{E}[X]) \neq 0$, Jensen's inequality must be strict. □

An immediate application is:

Corollary 15. $\mathrm{Var}(X) \geqslant 0$, *with equality if and only if* $\mathbb{P}(X = \mathbb{E}[X]) = 1$

Of course this follows from $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, but we can also do it with Jensen: $x \mapsto x^2$ is strictly convex, so $\mathbb{E}[X^2] \geqslant \mathbb{E}[X]^2$, and the result follows.

2.2 Finite applications

Most of the results in the sequel also have significant application in areas outside probability.

First, we look at some simple results we can derive using Jensen's inequality. While it appears that this is an easy way to derive lots of classical inequalities, see Remark 23 at the end for the bad news.

Theorem 16 (Finite weighted AM–GM–HM inequality). *Let* $(x_k)_{k=1}^n$ *be positive real numbers, and* $(\alpha_k)_{k=1}^n$ *satisfy* $w_k \in (0,1)$ *and* $\sum_{k=1}^n w_k = 1$. *Then*

$$\left(\sum_{k=1}^n w_k x_k^{-1}\right)^{-1} \leqslant \prod_{k=1}^n x_k^{w_k} \leqslant \sum_{k=1}^n w_k x_k,$$

with equality if and only if all x_k *are equal.*

Proof. Let X be a discrete random variable taking value x_k with probability w_k. The function $\varphi(x) = -\log x$ is strictly convex on $(0,\infty)$. Jensen then implies that $\mathbb{E}[-\log X] \geqslant -\log(\mathbb{E}[X])$, with equality when all the x_k are equal; since $\exp$ is strictly increasing, this is the same as saying

$$\mathbb{E}[X] \geqslant \exp(\mathbb{E}[\log X])$$

We calculate

$$\mathbb{E}[X] = \sum_{k=1}^n w_k x_k, \qquad \exp(\mathbb{E}[\log X]) = \exp\left(\sum_{k=1}^n w_k \log x_k\right) = \prod_{k=1}^n x_k^{w_k},$$

and the first inequality follows. For the second, the first inequality still holds (with the same equality conditions) if we replace x_k by $1/x_k$. This implies that

$$\sum_{k=1}^n w_k x_k^{-1} \geqslant \prod_{k=1}^n x_k^{-w_k}.$$

Since $x \mapsto 1/x$ is strictly decreasing, reciprocating gives the result. □

When the α_k are all equal, we have

Corollary 17 (Finite AM–GM–HM inequality[8]). *Let* $(x_k)_{k=1}^n$ *be positive real numbers. Then*

$$n\left(\sum_{k=1}^n x_k^{-1}\right)^{-1} \leqslant \left(\prod_{k=1}^n x_k\right)^{1/n} \leqslant \frac{1}{n}\sum_{k=1}^n x_k,$$

with equality if and only if all x_k *are equal.*

[8] The origins of this inequality are somewhat obscure. It was known to Gauss.

From now on, none of this material is in the schedule for Probability. It is included for reference and interest value.

We can use a simple instance of this to prove another very useful inequality:

Theorem 18 (Finite Cauchy–Schwarz inequality[9]). *Let Let $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ be complex numbers, and $(w_k)_{k=1}^n$ satisfy $w_k \in (0,1)$ and $\sum_{k=1}^n w_k = 1$. Then*

$$\sum_{k=1}^n |x_k||y_k|w_k \leqslant \left(\sum_{k=1}^n |x_k|^2 w_k\right)^{1/2}\left(\sum_{k=1}^n |y_k|^2 w_k\right)^{1/2}$$

Equality occurs if and only if there are constants λ, μ so that $\lambda|x_k| = \mu|y_k|$ for every k.

Proof. If one of the sums on the right is zero, every summand in that sum must be zero and the result is immediate since both sides are zero. Otherwise, dividing by the sums on the right, we have

$$\sum_{k=1}^n \frac{|x_k|}{(\sum_{i=1}^n |x_i|^2 w_i)^{1/2}} \frac{|y_k|}{(\sum_{i=1}^n |y_i|^2 w_i)^{1/2}} w_k.$$

Applying the AM–GM inequality to the product gives

$$\frac{|x_k|}{(\sum_{i=1}^n |x_i|^2 w_i)^{1/2}} \frac{|y_k|}{(\sum_{i=1}^n |y_i|^2 w_i)^{1/2}} \leqslant \frac{1}{2}\left(\frac{|x_k|^2}{\sum_{i=1}^n |x_i|^2 w_i} + \frac{|y_k|^2}{\sum_{i=1}^n |y_i|^2 w_i}\right),$$

with equality if and only if

$$|x_k|^2 \sum_{i=1}^n |y_i|^2 w_i = |y_k|^2 \sum_{i=1}^n |y_i|^2 w_i. \tag{2}$$

Multiplying this inequality by w_k and summing over k gives the result, with the equality case requiring that (2) is satisfied for each k. □

This has a significant generalisation:

Theorem 19 (Finite Hölder inequality[11]). *Let $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ be complex numbers, and $(\alpha_k)_{k=1}^n$ be positive. Let also $p, q > 1$ with $1/p + 1/q = 1$. Write $S(z) = \sum_{k=1}^n z_k \alpha_k$. Then*

$$|S(xy)| \leqslant S(|xy|) \leqslant S(|x|^p)^{1/p} S(|y|^q)^{1/q}.$$

Equality in the second occurs if and only if the sequences $(|x_k|^p)_{k=1}^n$ and $(|y_k|^q)_{k=1}^n$ are proportional.

One can generalise this further, to more than two sequences:

Theorem 20 (General finite Hölder inequality). *Let $(x_k^{(j)})_{k=1}^n$ be complex numbers, for $j \in \{1, \dots, m\}$, let $(\alpha_k)_{k=1}^n$ be positive. Let also $p_j > 1$ with $\sum_{j=1}^m p_j^{-1} = 1$. Write $S(z) = \sum_{k=1}^n z_k \alpha_k$. Then*

$$\left|S\left(\prod_{j=1}^m x^{(j)}\right)\right| \leqslant S\left(\prod_{j=1}^m |x^{(j)}|\right) \leqslant \prod_{j=1}^m \left(S(|x^{(j)}|^{p_j})\right)^{1/p_j}.$$

Equality in the second occurs if and only if all the sequences $(|x_k^{(j)}|^{p_j})_{k=1}^n$ are proportional.

[10] Rather delightfully, this paper uses $\mathcal{S}$es to denote integrals.

[9] The nomenclature here is famously incomplete. The unweighted version of this inequality was first proven by Cauchy [6, 7, Note II, Theorem XVI]. (Beware the page numbering, which differs between versions.). The version we discuss later for integrals was first proven by Bunyakovsky [5],[10] but is usually credited to Schwarz, whose proof was first published some years later [25]. Bunyakovsky also has the significant disadvantage of a wealth of available transliterations of his name.

[11] Why this inequality was attributed to Hölder is a mystery: not only was it first proven by Rogers [24] a year earlier, but Hölder [12] even cites it from Roger's paper! *caveat lector*, the inequality found in both of these papers is written in a less symmetrical form than the version that has become standard. The first statement and proof of the form given here seems to be by F. Riesz [22, § 33, p. 43]. See also [17] for more details.

This can be proved using the previous result and induction.

The other important inequality is a generalisation of the triangle inequality:

Theorem 21 (Finite Minkowski inequality [20, § 40, pp. 115–117]). *Let $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ be complex numbers, and $(\alpha_k)_{k=1}^n$ be positive. Let also $p > 1$, and again write $S(z) = \sum_{k=1}^n z_k \alpha_k$. Then*

$$S(|x+y|^p)^{1/p} \leqslant S(|x|^p)^{1/p} + S(|y|^p)^{1/p}.$$

Equality occurs if and only if the sequences $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ are proportional with nonnegative proportionality constants.

The following simple proof is due to F. Riesz [23].

Proof. We have

$$|x_k + y_k|^p = |x_k + y_k||x_k + y_k|^{p-1} \leqslant |x_k||x_k + y_k|^{p-1} + |y_k||x_k + y_k|^{p-1},$$

with equality if and only if x_k and y_k are proportional with a positive proportionality constant. Applying now the linear positive operator S to both sides,

$$S(|x+y|^p) \leqslant S\big(|x||x+y|^{p-1}\big) + S\big(|y||x+y|^{p-1}\big),$$

and now writing using Hölder's inequality on each term on the right,

$$S(|x+y|^p) \leqslant S(|x|^p)^{1/p} S(|x+y|^{(p-1)q})^{1/q} + S(|y|^p)^{1/p} S(|x+y|^{(p-1)q})^{1/q},$$

where as before $1/q = 1 - 1/p$, and equality holds if $|x|^p, |y|^p$ and $|x+y|^{(p-1)q}$ are all proportional. But $(p-1)q = p$, so in fact this equation is

$$S(|x+y|^p) \leqslant \Big(S(|x|^p)^{1/p} + S(|y|^p)^{1/p}\Big) S(|x+y|^p)^{1-1/p}$$

and dividing by $S(|x+y|^p)^{1-1/p}$ gives the result. □

Theorem 22 (General finite Minkowski inequality). *Let $(x_k^{(j)})_{k=1}^n$ be complex numbers, for $j \in \{1, \dots, m\}$, let $(\alpha_k)_{k=1}^n$ be positive, and $p > 1$. Then*

$$S\left(\sum_{j=1}^m |x^{(j)}|^p\right)^{1/p} \leqslant \sum_{j=1}^m \left(S(|x^{(j)}|^p)\right)^{1/p}.$$

Equality occurs if and only if all of the sequences $(x_k^j)_{k=1}^n$ are proportional with nonnegative proportionality constants.

This again follows from the previous result by induction.

Remark 23. All of the inequalities in this section could involve only rational numbers, in which case it seems more sensible to prove them using more elementary means than Jensen's inequality and the logarithm function, both of which currently rely on limits.[12] While this can be avoided for Jensen's inequality, at least for random variables with rational probabilities on finite sets (by induction), it cannot for the logarithm.

There are various methods to implement this, but the simplest is probably to begin by proving *Bernoulli's inequality*,[13]

$$x^r - 1 > r(x-1)$$

[12] Where this occurs in Jensen's inequality may not be obvious: remember that neither the tangent nor m in the support line are guaranteed to exist if we are only working in $\mathbb{Q}$.

[13] While Bernoulli did publish this inequality [2, Cap. IV, p. 380], it goes back at least to Sluse [26, Cap. IV, pp. 114–117]...

for $x \neq 1$ and $r > 1$ rational. From this we obtain the reverse inequality

$$x^r - 1 < r(x-1)$$

for $x \neq 1$ and $0 < r < 1$ rational, and then by substituting $x = a/b$ and multiplying by b,

$$a^r b^{1-r} - 1 < r(a-b) \implies a^r b^{1-r} < ra + (1-r)b,$$

which is the weighted AM–GM inequality for two numbers. If $\sum_{k=1}^{n} r_i = s = 1 - r_{n+1} < 1$ and we already know weighted AM–GM for n numbers, then

$$\begin{aligned}\prod_{k=1}^{n+1} x_k^{r_k} &= \left(\prod_{k=1}^{n} x_k^{r_k/s}\right)^s x_{k+1}^{1-s} \\ &< s\left(\prod_{k=1}^{n} x_k^{r_k/s}\right) + (1-s)x_{k+1} \\ &\leqslant s\sum_{k=1}^{n} \frac{r_k}{s} x_k + (1-s)x_{n+1} = \sum_{k=1}^{n+1} r_k x_k,\end{aligned}$$

with equality requiring that all x_k are equal. Hence the result follows by induction.

Our proofs of the other inequalities in this section do not require modification, since they do not use any results true of the real numbers but not the rationals.

For more details of the development along these lines, one can see, e.g. [10, § 74, pp. 143f, Appendix I, pp. 487–491], [9] or [11, § 2.14–15, pp. 37–42].

2.3 Continuous applications

Everything in the previous section has an integral analogue. These are best derived straight from Jensen's inequality, rather than by using a limiting argument, so as to avoid destroying strict inequality wherever possible. However, the conditions for equality are no longer so simple, and we shall omit them.

Another pitfall of continuous random variables is that either side of the inequality could be infinite. The interpretation for this should be that if the smaller side is infinite, the larger side must be too.

Theorem 24 (General AM–GM for random variables). *For any random variable X with $\mathbb{P}(X > 0) = 1$,*

$$\mathbb{E}\left[\frac{1}{X}\right] \leqslant \exp(\mathbb{E}[\log X]) \leqslant \mathbb{E}[X]$$

Theorem 25. *For any random variable X, if $q < p$,*

$$(\mathbb{E}[|X|^q])^{1/q} \leqslant (\mathbb{E}[|X|^p])^{1/p}.$$

Proof. Apply Jensen to the random variable X^q and the convex function $x \mapsto x^{p/q}$. □

The proofs of the following are very close to their discrete counterparts and so are omitted.

Theorem 26 (Hölder's inequality for random variables). *Let $(X_j)_{j=1}^m$ be complex-valued random variables. If $p_j > 1$ with $\sum_{j=1}^m 1/p_j = 1$, then*

$$\left|\mathbb{E}\left[\prod_{j=1}^{m} X_j\right]\right| \leqslant \mathbb{E}\left[\prod_{j=1}^{m} |X_j|\right] \leqslant \prod_{j=1}^{m} (\mathbb{E}[|X|^{p_j}])^{1/p_j}.$$

Theorem 27 (Minkowski's inequality for random variables). *Let $(X_j)_{j=1}^m$ be complex-valued random variables. If $p > 1$, then*

$$\left(\mathbb{E}\left[\left|\sum_{j=1}^m X_j\right|^p\right]\right)^{1/p} \leqslant \sum_{j=1}^m (\mathbb{E}[|X|^p])^{1/p}.$$

2.4 Integral inequalities

These can all also be expressed for functions: all integrals in the following are over $\mathbb{R}$.

Theorem 28 (AM–GM for integrals). *Let $f : \mathbb{R} \to (0, \infty)$, and let $w : \mathbb{R} \to [0, \infty)$ with $\int w = 1$. Then*

$$\int \frac{w}{f} \leqslant \exp\left(\int w \log f\right) \leqslant \int fw.$$

Theorem 29. *Let $f : \mathbb{R} \to \mathbb{C}$, and let $w : \mathbb{R} \to [0, \infty)$ with $\int w = 1$. Then if $q < p$,*

$$\left(\int |f|^q\right)^{1/q} \leqslant \left(\int |f|^p\right)^{1/p}$$

Theorem 30 (Hölder's inequality for integrals). *For $j \in \{1, \dots, m\}$, let $f_j : \mathbb{R} \to \mathbb{C}$, and $w : \mathbb{R} \to [0, \infty)$. Then if $1 < p_j$ with $\sum_{j=1}^m 1/p_j = 1$,*

$$\left|\int \left(\prod_{j=1}^m f_j\right) w\right| \leqslant \int \left(\prod_{j=1}^m |f_j|\right) w \leqslant \prod_{j=1}^m \left(\int |f_j|^{p_j} w\right)^{1/p_j}.$$

The best-known special case of this is

Theorem 31 (Schwarz's inequality[14]). *Let $f, g : \mathbb{R} \rightrightarrows \mathbb{C}$, and $\alpha : \mathbb{R} \to [0, \infty)$. Then*

$$\left|\int fg\alpha\right| \leqslant \int |f||g|\alpha \leqslant \left(\int |f|^2\alpha\right)^{1/2} \left(\int |g|^2\alpha\right)^{1/2}.$$

Finally,

Theorem 32 (Minkowski's inequality for integrals). *For $j \in \{1, \dots, m\}$, let $f_j : \mathbb{R} \to \mathbb{C}$, and $w : \mathbb{R} \to [0, \infty)$. Then if $p > 1$,*

$$\left(\int \left|\sum_{j=1}^m f_j\right|^p w\right)^{1/p} \leqslant \sum_{j=1}^m \left(\int |f_j|^p w\right)^{1/p}. \tag{3}$$

Remark 33. It should be emphasised that inequalities like Jensen's and AM–GM are of a different character from inequalities like Hölder's and Minkowski's, in that the former are not *homogeneous*: it is a requirement that the w_i sum to 1 (i.e. that they form a probability measure) in the former, but not in the latter; this is because if we rescale w to λw, the powers of λ on both sides of the latter are the same, but this does not make sense for the former. Similarly, the functions in the latter can be rescaled in the same way.

Remark 34. With a sufficiently general definition of integral,[15] these inequalities all generalise to integrals on *any* measure space.

[14] Actually Bunyakovsky's. See footnote on Theorem 18.
[15] Namely the Lebesgue integral.

3 Some further inequalities

We define the *p-norm*: for $w \geqslant 0$,

$$\|f\|_p = \left(\int |f|^p w \right)^{1/p}.$$

With this, we can express Hölder's inequality as

$$\|fg\|_1 \leqslant \|f\|_p \|g\|_q$$

and Minkowski's inequality as

$$\|f + g\|_p \leqslant \|f\|_p + \|g\|_p,$$

which is the triangle inequality for the p-norm, necessary for proving that the p-norm is a norm. Judicious application of Hölder gives us still more inequalities:

Besso's inequality [16] If $q < p$, and $\int w = 1$,

$$\|f\|_q \leqslant \|f\|_p.$$

(The condition on w is required because this inequality is not homogeneous.) Hence the p-norm is a nondecreasing function of p.

Littlewood's inequality [16, Lemma 2, p. 169] With $\frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$,

$$\|f\|_{p_\theta} \leqslant \|f\|_{p_0}^{1-\theta} \|f\|_{p_1}^{\theta}.$$

Lyapunov's inequality [15, p. 2] With $p = (1-\theta)p_0 + \theta p_1$,

$$\|f\|_p^p \leqslant \|f\|_{p_0}^{p_0(1-\theta)} \|f\|_{p_1}^{p_1\theta}.$$

These last two are *interpolation inequalities*: they bound intermediate norms. In particular, this implies that the functions $\|f\|_p$ and $\|f\|_p^p$ have convex logarithms.

The following is a significant generalisation of Minkowski's inequality:

Theorem 35. *Let $f : X \times Y \to \mathbb{C}$ and $p > 1$. Then*

$$\left(\int_X \left| \int_Y f(x,y)\, dv(y) \right|^p d\mu(x) \right)^{1/p} \leqslant \int_Y \left(\int_X |f(x,y)|^p \, d\mu(x) \right)^{1/p} dv(y),$$

with equality if and only if $f(x,y) = g(x)h(y)$ for some g and h.

The proof can be done by "integrating Hölder's inequality".

Remark 36. As usual, the nomenclature is somewhat suspect here: this is usually called "Minkowski's integral inequality" or "Minkowski's inequality for integrals", but these suffer from two problems: firstly, this inequality is not due to Minkowski, being rather more general than what he proved, and secondly, either of these names could equally apply to (3). Therefore a different name is required. According to [11, p. 31], this was known to Ingham before 1929, and published by Jessen in 1931. Therefore the obvious candidates are "Ingham–Jessen inequality", "Minkowski–Jessen inequality", or, if the reader really wants to use ink, "Minkowski–Ingham–Jessen inequality".[17]

[16] Supposely this was first shown by Besso, but we have been unable to find a copy of the paper [4]. This inequality is sometimes incorrectly attributed to Lyapunov.

[17] As we shall see next, "Jessen's inequality" is taken.

Finally, we have an extension to more general operators than the expectation. Let X be nonempty, L be a vector space of functions $X \to \mathbb{R}$ that contains the constant function 1. A linear functional $A\colon L \to \mathbb{R}$ is called *isotonic* if

$$(\forall f \in L) \qquad f \geqslant 0 \implies A(f) \geqslant 0;$$

of course linearity implies that for $f \geqslant g$, $A(f) \geqslant A(g)$. It is extremely easy to come up with such functionals:

1. $X \mapsto \mathbb{E}[X]$
2. $(a_k)_{k=1}^n \mapsto \sum_{k=1}^n a_k w_k$, where $w_k \geqslant 0$
3. $f \mapsto \int_X f \, d\mu$, where μ is a measure on X.

These are all the cases we have considered so far. Now, the unifying result is:

Theorem 37 (Jessen's inequality [14]). *Let L and A be as above, with $0 < A(1) < \infty$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Then for any $f \in L$ with $\varphi(f) \in L$,*

$$\varphi\left(\frac{A(f)}{A(1)}\right) \leqslant \frac{A(\varphi(f))}{A(1)}.$$

The proof is in fact very similar to the proof of Jensen's inequality. This is not a coincidence: the whole point of the conditions on the functions is to make the previous proof work! If you really want to see a proof, Jessen's original article or Beesack and Pečarić's more recent article [3] on related inequalities will do. This inequality in turn gives rise to A-equivalents of AM–GM, Hölder, Minkowski and so on.

4 Further reading

The study of inequalities in their own right, rather than as a means to an end, was initiated by Hardy, Littlewood and Pólya's book *Inequalities* [11]. The modern classic of the genre is widely regarded as being Steele's *The Cauchy–Schwarz Master Class* [27].

References

[1] Bather, J. 'A conversation with Herman Chernoff'. *Statistical Science* 11.4 (Nov. 1996), pp. 335–350.

[2] Bernoulli, J. *Positiones Arithmeticae de Seriebus Infinitis*. Basel, 1689.

[3] Bessack, P. R. and Pečarić, J. E. 'On Jessen's Inequality for Convex Functions'. *Journal of Mathematical Analysis and Applications* 110 (1985), pp. 536–552.

[4] Besso, D. 'Teoremi elementari sui massimi e minimi'. *Annuario de Ist. Tecnico di Roma* (1879), pp. 7–24. Unable to locate a copy. Reportedly reprinted in Boll. di Mat. 8 (1907).

[5] Bouniakowsky, V. 'Sur quelques inegalités concernant les intégrales aux différences finies'. *Mémoires de L'Académie Impériale des Sciences de St.-Pétersbourg, VII^e Série* 1.9 (1859), pp. 1–18.

[6] Cauchy, A.-L. *Cours d'analyse de l'École royale polytechnique*. L'Imprimerie Royale, 1821. Rpt. as *Œuvres Completes*. Vol. 2.3. Gauthier-Villars, 1897.

[7] Cauchy, A.-L. *Cauchy's Cours d'Analyse: An Annotated Translation*. Trans. by Bradley, R. E. and Sandifer, C. E. Springer, 2009. ISBN: 978-1-4419-0548-2. Translation of [6].

[8] Chernoff, H. 'A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations'. *The Annals of Mathematical Statistics* 23.4 (Dec. 1952), pp. 493–507.

[9] Hardy, G. H. 'Prolegomena to a Chapter on Inequalities'. *Journal of the London Mathematical Society* s1-4.1 (1929), pp. 61–78.

[10] Hardy, G. H. *A course of pure mathematics*. With a forew. by Körner, T. W. 10th Centenary Edition. Cambridge University Press, 2008. xix+509. ISBN: 9780521720557.

[11] Hardy, G. H. and Littlewood, J. E. and Pólya, G. *Inequalities*. 2nd ed. Cambridge University Press, 1952. xii+324. Reprinted 1959, 1964, 1967, 1973, 1988 (twice). First paperback edition 1988; reprinted 1989, 1991, 1994, 1997, 1999. Transferred to digital printing 2001.

[12] Hölder, O. 'Ueber einen Mittelwerthsatz'. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen* 2 (1889), pp. 38–46.

[13] Jensen, J. L. W. V. 'Sur les fonctions convexes et les inégalités entre les valeurs moyennes'. *Acta Mathematica* 30 (1906), pp. 175–193.

[14] Jessen, B. 'Bemærkninger om konvekse Funktioner og Uligheder imellem Middelværdier, I'. *Matematisk Tidsskrift B* (1931), pp. 17–28.

[15] Liapounoff, A. 'Nouvelle forme du théorème sur la limite de probabilité'. *Mémoires de L'Académie Impériale des Sciences de St.-Pétersbourg, VII^e Série* 12.5 (1901).

[16] Littlewood, J. E. 'On bounded bilinear forms in an infinite number of variables'. *The Quarterly Journal of Mathematics* os-1.1 (Jan. 1930), pp. 164–174. ISSN: 0033-5606. eprint: https://academic.oup.com/qjmath/article-pdf/os-1/1/164/4482719/os-1-1-164.pdf.

[17] Maligranda, L. 'Why Hölder's inequality should be called Rogers' inequality'. *Mathematical Inequalities & Applications* 1.1 (1998), pp. 69–83.

[18] Markoff, A. A. *Wahrscheinlichkeitsrechnung*. B.G. Taubner, 1912.

[19] Markov, A. A. *Ischislenie Veroiatnostei (Calculus of Probabilities)*. St Petersburg: Imperial Academy of Sciences, 1900.

[20] Minkowski, H. *Geometrie der Zahlen*. B.G. Taubner, 1910.

[21] Philips, T. K. and Nelson, R. 'The Moment Bound Is Tighter Than Chernoff's Bound for Positive Tail Probabilities'. *The American Statistician* 49.2 (1995), pp. 175–178.

[22] Riesz, F. *Les systèmes d'équations linéaires à une infinité d'inconnues*. Gauthier-Villars, 1913.

[23] Riesz, F. 'Su alcune disuguaglianze.' Italian. *Bollettino della Unione Matematica Italiana* 7 (1928), pp. 77–79. ISSN: 0041-7084.

[24] Rogers, L. J. 'An extension of a certain theorem in inequalities'. *Messenger of Mathematics* 7 (1888), pp. 145–150.

[25] Schwarz, H. A. 'Über ein Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung'. *Acta Societatis Scientiarum Fennicae* XV (1888), pp. 318–362.

[26] Sluse, R. F. *Mesolabum*. Streel, 1668.

[27] Steele, J. M. *The Cauchy–Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004. ISBN: 9780521546775.

[28] Tchébychef, P. 'Des valeurs moyennes'. Trans. by Khanikof, N. de. *Journal de Mathématiques Pures et Appliquées, Série 2* 12 (1867), pp. 177–184.