Case study: The arithmetic-geometric means inequality

Mathematical results are not just inert facts, but can live in a variety of different "neighbourhoods". This is not generalluy evident from the mere statement of the result, but is likely to be seen from the proof. To prove a result is to first adopt a point of view towards the result and to put it into some context, from which the strategy of the proof will be constructed.

A nice example of this is the arithmetic-geometry means inequality. We will discuss various arguments for this and what they convey to us about the result. Recall that the inequality is

$$\sqrt{ab} \le \frac{1}{2}(a+b) \tag{1}$$

whenever a and b are nonnegative real numbers.

1. *Meaning and interpretation.* The inequality (1) is given in algebraic terms. However, putting side the case where one of a and b might vanish (which is obvious), we can express in a more informal way: *the geometric mean of two positive numbers does not exceed their arithemtic mean.* The geometric mean g of a and b is that positive number g for which $a/g = g/b$; the arithmetic mean is that number m for which $a - m = m - b$. Then we have to argue that $g \le m$.

We can view the situation in this way. Without losing generality, let us suppose that $a \ge b$. The condition for the geometric mean means that the factor that we multiply a by to get g is the same as the factor we multiply g by to get b. Since a is larger than g, this means that we have to "come down" further from a to get to g than we have to go from g to b. Thus, going down to g takes us more than halfway between a and b, and so below m, the exact halfway point.

The foregoing intuitive and informal explanation may leave much to be desired from the point of view of a properly composed proof, but it exploits the semantic content of the terms involved in a way that the more formal arguments given later fail to do. We tackle the inequality at the level of its meaning, and the role of this argument is to bring to the front of our minds the essence of the result.

2. *The straightforward technical argument.* A general strategy for establishing inequalities is to look at the difference of the members on both sides of the inequality and thus convert it to a relationship between some quantity and 0. In order to avoid the

awkwardness of the surd, we establish the equivalent inequaliy

$$4ab \leq (a+b)^2 \ .$$

Subtracting the left member from the right, we obtain the quantity

$$(a+b)^2 - 4ab = a^2 - 2ab + b^2 = (a-b)^2 \tag{2}$$

which is always greater than or equal to 0, with equality occurring only if $a = b$. ♠

While the details are specific to the inequality, the strategy is a general one and places the inequality is a class of similar polynomial inequalities. For example, the generalization of this to three numbers is the following:

$$\sqrt[3]{abc} \leq \frac{1}{3}(a+b+c) \ ,$$

for nonnegative values of a, b, c. The substitutions $x = u^3$, $y = v^3$, $z = w^3$ yields the equivalent

$$3uvw \leq u^3 + v^3 + w^3 \ .$$

The strategy works on this; just note that

$$\begin{aligned} u^3 + v^3 + w^3 - 3uvw &= (u+v+w)(u^2+v^2+w^2-uv-vw-wu) \\ &= \frac{1}{2}(u+v+w)[(u-v)^2+(v-w)^2+(w-u)^2] \geq 0 \ . \end{aligned}$$

At this stage, the proof generates some questions for investigation. The analogue of these inequalities for n variables amounts to

$$nu_1u_2\cdots u_n \leq u_1^n + u_2^n + \cdots + u_n^n$$

for nonnegative values of the u_i, and we might ask whether a similar strategy works to establish this, *viz*, to write the difference

$$u_1^n + \cdots + u_n^n - nu_1u_2\cdots u_n$$

as a product of polynomials each of which is clearly seem to be nonnegative for some elementary reason.

Observe that this argument is a contrast to the first. While the first brings out the interpretation of the means and their relationship, this one is completely technical and so becomes an exemplar for what might be done in similar situations. It does not require the understanding of the situation that the first argument evokes.

3. *The theory of the quadratic.* The proof that one might adduce for a result depends very much on what one sees as the significant aspects of the situation. Argument 1 saw that significance in the origins of the means an Argument 2 saw the possibility of taking a difference that could be shown as positive. Here we observe that there is an interplay between the sum and the product of two quantitites, which reminds us of a quadratic polynomial whose coefficients can be described in terms of the sum and product of their roots. Indeed, consider the quadratic equation:

$$0 = (x - a)(x - b) = x^2 - (a + b)x + ab\ .$$

It (clearly) has real roots, and so its discriminant must be nonnegative:

$$(a + b)^2 - 4ab \geq 0\ .$$

But, of course, this is precisely what we are looking for!

4. *A geometric model.* How we see a mathematical result is reflected in how we model it in our minds, and this can be evoked in the proof. For example, consider a circle with diameter AB of length $a + b$. Determine the point P on the diameter for which $AP = a$ and $PB = b$, and draw a perpendicular to AB that meets the circumference of the circle at Q. Appealing to the similar triangles APQ and QPB, we see that $A : PQ = PQ : QB$ whereupon $PQ - \sqrt{ab}$. But PQ is no greater than the radius of the circle (in particular, that radius to which it is parallel), and the result follows.

5. *Concavity.* Anyone with a structural understanding of the arithmetic operations will observe that the geometric and arithmetic means are analogues with multiplication and exponentiation in the former playing the roles of addition and multiplication in the latter The conduit of this relationship is the logarithm function. Accordingly, in the case that a and b are both positive, we see that the arithmetic-geometric means inequality is equivalent to

$$\frac{1}{2}(\log a + \log b) \leq \log(\frac{1}{2})(a + b)\ .$$

If we know, for some reason, that the logarithm function is concave (*i.e.* that if satisfies the condition

$$f(ta + (1 - t)b) \geq tf(a) + (1 - t)f(b)$$

whenever $0 \leq t \leq 1$), then this inequality, being a special case, is immediate.

The validity of this proof depends on what we are permitted to assume at this stage. If we need the arithmetic-geometric means inequality to establish the concavity of the logarithm, then the argument is circular, and all we have done is to state a connection between two properties that may or may not be true. However, if we can establish independently the concavity of the logarithm function (say by noting that its second derivative is negative), then we not only have a neat structural argument for the inequality but, returning by the route by which we came, a bounty of additional results:

$$a^t b^{1-t} \leq ta + (1-t)b$$

whenever $a, b > 0$ and $0 \leq t \leq 1$. Because this holds for arbitrary exponents t in the closed unit interval, we have gotten a result that is beyond the realm of algebra and is the fruit of analysis.

This argument opens up other possible areas of investigation. The logarithm function is not the only concave function; they abound. Each such function generates its own inequality. Some of these may be mundane and devoid of interest or application, but others may have wide-ranging significance.

6. *An abstract approach.* While the foregoing argument leads us to a level of generality, a more acute use of the logarithm function and of the perception that the logarithm of the geometric means is the arithmetic mean of the logarithms leads us to an argument of astonishing simplicity and generality. The price we pay for this is an abstractness, which may drain the result of some of its immediate semantic content but crystallizes its structural aspects.

Let S be an arbitrary set an let $\mathbf{R}$ denote the system of real numbers. For each function f from S to $\mathbf{R}$, we suppose that an *average* $A(f)$ is defined that satisfies these five axioms:

(i) $A(f)$ is a real number;

(ii) $A(cf) = cA(f)$ fpr amy real constant c;

(iii) $A(f+g) = A(f) + A(g)$ for any two functions f and g;

(iv) $A(f) \geq 0$ for all functions f what assume only nonnegative values;

(v) $A(\mathbf{1}) = 1$, where $\mathbf{1}$ denotes the function on s that everywhere assumes the value 1.

In other words, A is a positive normalized linear functional on the vector space of real functions from S to $\mathbf{R}$. Let us call $A(f)$ the *arithmetic mean* of f. It is a consequence of (iii) and (iv) that

(vi) $A(f) \leq A(g)$ whenever $f \leq g$ (*i.e.* $f(s) \leq g(s)$ for each $s \in S$. This is the condition of *monotonicity*.

If f is a function on S that is everywhere positive, then $\log f$, the natural logarithm of f is everywhere defined. We define the *geometric mean* $G(f)$ by

$$G(f) = \exp A(\log f)$$

where $\exp x = e^x$, e being the base of natural logarithms.

We wish to establish that

$$G(f) \leq A(f) \tag{4}$$

whenever f is a strictly positive function defined on S. To avoid needless complications, we will confine our considerations to the situation where $A(f) > 0$.

For example, if we let S be a set with exactly two points, then each real function on S can be represented by a pair (a, b) of real numbers, and we can take $A(a, b) = \frac{1}{2}(a + b)$.

It is a consequence of the concavity of the logarithm function as well as a simple observation from the graph of the logarithm that $\log t \leq t - 1$ for $t > 0$. Let s be any point of S and apply this inequality to $t = f(s)/A(f)$. Then, by monotonicity,

$$\log \left(\frac{f}{A(f)} \leq \frac{f}{A(f)} - \mathbf{1}\right.$$

since this holds whenever both sides are evaluated at any particular value of s.

From the characteristic property of the logarithm, this inequality can be converted to

$$\log f - \log A(f)\mathbf{1} \leq \frac{f}{A(f)} - \mathbf{1}\ . \tag{5}$$

(Recall that $A(f)$ is a real number. We regard this as an inequality between two functions.

Taking the average of the left side and applying properties (ii), (iii) and (v) of A, we find that

$$\begin{aligned} &A(\log f - \log A(f)\mathbf{1}) \\ &\qquad = A(\log f) - A(\log A(f)\mathbf{1} \\ &\qquad = A(\log f) - (\log A(f))A(\mathbf{1}) = A(\log f) - (\log A(f))\ . \end{aligned}$$

Taking the average of the right side and applying properties (ii) and (v), we get that

$$A(f/A(f) - \mathbf{1}) = A(f/A(f)) - A(\mathbf{1}) = (1/A(f))A(f) - 1 = 1 - 1 = 0 \ .$$

Thus, (5) yields $A(\log f) - (\log A(f)) \leq 0$, or

$$A(\log f) \leq \log A(f) \ .$$

Exponentiating both sides of this inequality leads to the desired $G(f) \geq A(f)$. ♠

This argument transports us to a different world, and may be difficult for someone to follow who has not already had experience in the more abstract realms of mathematics. At each stage, one may have to consciously reflect on exactly what kind of quantity each of the symbols represents and in order to ascertain that the operations carried out are properly sited. The arithmetic-geometric means inequality that we started out with is a picayune application of this result. More generally, we can apply it to an operator defined on a space of real functions defined on a set of n points; such a function can be represented by a vector $f = (a_1, a_2, \cdots, a_n)$ displaying its respective values.

Consider a vector $(w_1, w_2, \cdots, w_n)$ of *weights* that satisfy the two conditions that each w_i is nonnegative and that $w_1 + w_2 + \cdots + w_n = 1$. Then we can define the weighted average

$$A(f) = w_1 a_1 + w_2 a_2 + \cdots + w_n a_n \ .$$

The general result that has just been obtained yields

$$a_1^{w_1} a_2^{w_2} \cdots a_n^{w_n} \leq w_1 a_1 + w_2 a_2 + \cdots + w_n a_n \ .$$

We have here six different arguments of the arithmetic-geometric means inequality, and each of them conveys a different flavour of the inequality. Some of them reach out to related areas of mathematics, enriching these other areas with an increased sense of relevance, generality and value, while others plough into the very meaning that underlies the symbols.