

## BIG OMICRON AND BIG OMEGA AND BIG THETA

Donald E. Knuth  
Computer Science Department  
Stanford University  
Stanford, California 94305

Most of us have gotten accustomed to the idea of using the notation  $O(f(n))$  to stand for any function whose magnitude is upper-bounded by a constant times  $f(n)$ , for all large  $n$ . Sometimes we also need a corresponding notation for lower-bounded functions, i.e., those functions which are at least as large as a constant times  $f(n)$  for all large  $n$ . Unfortunately, people have occasionally been using the  $O$ -notation for lower bounds, for example when they reject a particular sorting method "because its running time is  $O(n^2)$ ." I have seen instances of this in print quite often, and finally it has prompted me to sit down and write a Letter to the Editor about the situation.

The classical literature does have a notation for functions that are bounded below, namely  $\Omega(f(n))$ . The most prominent appearance of this notation is in Titchmarsh's magnum opus on Riemann's zeta function [8], where he defines  $\Omega(f(n))$  on p. 152 and devotes his entire Chapter 8 to " $\Omega$ -theorems". See also Karl Prachar's Primzahlverteilung [7], p. 245.

The  $\Omega$  notation has not become very common, although I have noticed its use in a few places, most recently in some Russian publications I consulted about the theory of equidistributed sequences. Once I had suggested to someone in a letter that he use  $\Omega$ -notation "since it had been used by number theorists for years"; but later, when challenged to show explicit references, I spent a surprisingly fruitless hour searching in the library without being able to turn up a single reference. I have recently asked several prominent mathematicians if they knew what  $\Omega(n^2)$  meant, and more than half of them had never seen the notation before.

Before writing this letter, I decided to search more carefully, and to study the history of  $O$ -notation and  $o$ -notation as well. Cajori's two-volume work on history of mathematical notations does not mention any of these. While looking for definitions of  $\Omega$  I came across dozens of books from the early part of this century which defined  $O$  and  $o$  but not  $\Omega$ .

I found Landau's remark [6, p. 883] that the first appearance of  $O$  known to him was in Bachmann's 1894 book [1, p. 401]. In the same place, Landau said that he had personally invented the  $o$ -notation while writing his handbook about the distribution of primes; his original discussion of  $O$  and  $o$  is in [6, pp. 59-63].

I could not find any appearances of  $\Omega$ -notation in Landau's publications; this was confirmed later when I discussed the question with George Pólya, who told me that he was a student of Landau's and was quite familiar with his writings. Pólya knew what  $\Omega$ -notation meant, but never had used it in his own work. (Like teacher, like pupil, he said.)

Since  $\Omega$  notation is so rarely used, my first three trips to the library bore little fruit, but on my fourth visit I was finally able to pinpoint its probable origin: Hardy and Littlewood introduced  $\Omega$  in their classic 1914 memoir [4, p. 225], calling it a "new" notation. They used it also in their major paper on distribution of primes [5, see pp. 125ff], but they apparently found little subsequent need for it in later works.

Unfortunately, Hardy and Littlewood didn't define  $\Omega(f(n))$  as I wanted them to; their definition was a negation of  $o(f(n))$ , namely a function whose absolute value exceeds  $Cf(n)$  for infinitely many  $n$ , when  $C$  is a sufficiently small positive constant. For all the applications I have seen so far in computer science, a stronger requirement (replacing "infinitely many  $n$ " by "all large  $n$ ") is much more appropriate.

After discussing this problem with people for several years, I have come to the conclusion that the following definitions will prove to be most useful for computer scientists:

$O(f(n))$  denotes the set of all  $g(n)$  such that there exist positive constants  $C$  and  $n_0$  with  $|g(n)| \leq Cf(n)$  for all  $n \geq n_0$ .

$\Omega(f(n))$  denotes the set of all  $g(n)$  such that there exist positive constants  $C$  and  $n_0$  with  $g(n) \geq Cf(n)$  for all  $n \geq n_0$ .

$\Theta(f(n))$  denotes the set of all  $g(n)$  such that there exist positive constants  $C, C'$ , and  $n_0$  with  $Cf(n) \leq g(n) \leq C'f(n)$  for all  $n \geq n_0$ .

Verbally,  $O(f(n))$  can be read as "order at most  $f(n)$ ";  $\Omega(f(n))$  as "order at least  $f(n)$ ";  $\Theta(f(n))$  as "order exactly  $f(n)$ ". Of course, these definitions apply only to behavior as  $n \rightarrow \infty$ ; when dealing with  $f(x)$  as  $x \rightarrow 0$  we would substitute a neighborhood of zero for the neighborhood of infinity, i.e.,  $|x| \leq x_0$  instead of  $n \geq n_0$ .

Although I have changed Hardy and Littlewood's definition of  $\Omega$ , I feel justified in doing so because their definition is by no means in wide use, and because there are other ways to say what they want to say in the comparatively rare cases when their definition applies. I like the mnemonic appearance of  $\Omega$  by analogy with  $O$ , and it is easy to typeset. Furthermore, these two notations as defined above are nicely complemented by the  $\Theta$ -notation, which was suggested to me independently by Bob Tarjan and by Mike Paterson.

The definitions above refer to "the set of all  $g(n)$  such that ...", rather than to "an arbitrary function  $g(n)$  with the property that ..."; I believe that this definition in terms of sets, which was suggested to me many years ago by Ron Rivest as an improvement over the definition in the first printing of my volume 1, is the best way to define  $O$ -notation. Under this interpretation, when the  $O$ -notation and its relatives are used in formulas, we are actually speaking about sets of functions rather than single functions. When  $A$  and  $B$  are sets of functions,  $A+B$  denotes the set  $\{a+b \mid a \in A \text{ and } b \in B\}$ , etc.; and " $1+O(n^{-1})$ " can be taken to mean the set of all functions of the form  $1+g(n)$ , where  $|g(n)| \leq Cn^{-1}$ , for some  $C$  and all large  $n$ . The phenomenon of one-way equalities arises in this connection, i.e., we write  $1+O(n^{-1}) = O(1)$  but not  $O(1) = 1+O(n^{-1})$ . The equal sign here really means  $\subseteq$  (set inclusion), and this has bothered many people who propose that we not be allowed to use the  $=$  sign in this context. My feeling is that we should continue to use one-way equality together with  $O$ -notations, since it has been common practice of thousands of mathematicians for so many years now, and since we understand the meaning of our existing notation sufficiently well.

We could also define  $\omega(f(n))$  as the set of all functions whose ratio to  $f(n)$  is unbounded, by analogy to  $o(f(n))$ . Personally I have felt little need for these  $o$ -notations; on the contrary, I have found it a good discipline to obtain  $O$ -estimates at all times, since it has taught me about more powerful mathematical methods. However, I expect someday I may

have to break down and use  $o$ -notation when faced with a function for which I can't prove anything stronger.

Note that there is a slight lack of symmetry in the above definitions of  $O$ ,  $\Omega$ , and  $\Theta$ , since absolute value signs are used on  $g(n)$  only in the case of  $O$ . This is not really an anomaly, since  $O$  refers to a neighborhood of zero while  $\Omega$  refers to a neighborhood of infinity. (Hardy's book on divergent series uses  $O_L$  and  $O_R$  when a one-sided  $O$ -result is needed. Hardy and Littlewood [5] used  $\Omega_L$  and  $\Omega_R$  for functions respectively  $< -Cf(n)$  and  $> Cf(n)$  infinitely often. Neither of these has become widespread.)

The above notations are intended to be useful in the vast majority of applications, but they are not intended to meet all conceivable needs. For example, if you are dealing with a function like  $(\log \log n)^{\cos n}$  you might want a notation for "all functions which oscillate between  $\log \log n$  and  $1/\log \log n$  where these limits are best possible". In such a case, a local notation for the purpose, confined to the pages of whatever paper you are writing at the time, should suffice; it isn't necessary to worry about standard notations for a concept unless that concept arises frequently.

I would like to close this letter by discussing a competing way to denote the order of function growth. My library research turned up the surprising fact that this alternative approach actually antedates the  $O$ -notation itself. Paul du Bois-Reymond [2] used the relational notations

$$g(n) < f(n) \quad , \quad f(n) > g(n)$$

already in 1871, for positive functions  $f(n)$  and  $g(n)$ , with the meaning we can now describe as  $g(n) = o(f(n))$  (or as  $f(n) = \omega(g(n))$ ). Hardy's interesting tract on "Orders of Infinity" [3] extends this by using also the relations

$$g(n) \leq f(n) \quad , \quad f(n) \geq g(n)$$

to mean  $g(n) = O(f(n))$  (or, equivalently,  $f(n) = \Omega(g(n))$ ), since we are assuming that  $f$  and  $g$  are positive). Hardy also wrote

$$f(n) \asymp g(n)$$

when  $g(n) = \Theta(f(n))$ , and

$$f(n) \asymp g(n)$$

when  $\lim_{n \rightarrow \infty} f(n)/g(n)$  exists and is neither 0 nor  $\infty$ ; and he wrote

$$f(n) \sim g(n)$$

when  $\lim_{n \rightarrow \infty} f(n)/g(n) = 1$ . (Hardy's  $\asymp$  notation may seem peculiar at first, until you realize what he did with it; for example, he proved the following nice theorem: "If  $f(n)$  and  $g(n)$  are any functions built up recursively from the ordinary arithmetic operations and the  $\exp$  and  $\log$  functions, we have exactly one of the three relations  $f(n) < g(n)$ ,  $f(n) \asymp g(n)$ , or  $f(n) > g(n)$ .")

Hardy's excellent notation has become somewhat distorted over the years. For example, Vinogradov [9] writes  $f(n) \ll g(n)$  instead of  $f(n) \leq g(n)$ ; thus, Vinogradov is comfortable with the formula

$$200^2 \ll \binom{n}{2},$$

while I am not. In any event, such relational notations have intuitively clear transitive properties, and they avoid the use of one-way equalities which bother some people. Why, then, should they not replace  $O$  and the new symbols  $\Omega$  and  $\Theta$ ?

The main reason why  $O$  is so handy is that we can use it right in the middle of formulas (and in the middle of English sentences, and in tables which show the running times for a family of related algorithms, etc.). The relational notations require us to transpose everything but the function we are estimating to one side of an equation. (Cf. [7], p. 191.) Simple derivations like

$$\begin{aligned} \left(1 + \frac{H_n}{n}\right)^{H_n} &= \exp(H_n \ln(1 + H_n/n)) \\ &= \exp(H_n(H_n/n + O(\log n/n)^2)) \\ &= \exp(H_n^2/n + O((\log n)^3/n^2)) \\ &= \exp((\ln n + \gamma)^2/n + O((\log n)^3/n^2)) \\ &= (1 + O((\log n)^3/n^2))e^{(\ln n + \gamma)^2/n} \end{aligned}$$

would be extremely cumbersome in relational notation.

When I am working on a problem, my scratch paper notes often contain ad-hoc notations, and I have been using an expression like " $(\leq 5n^2)$ " to stand for the set of all functions which are  $\leq 5n^2$ . Similarly, I can write " $(\sim 5n^2)$ " to stand for functions which are asymptotic to  $5n^2$ , etc.; and " $(\leq n^2)$ " would therefore be equivalent to  $O(n^2)$ , if I made appropriate extensions of the  $\leq$  relation to functions which may be negative. This would provide a uniform notational convention for all sorts of things, for use in the middle of expressions, giving more than just the  $O$  and  $\Omega$  and  $\Theta$  proposed above.

In spite of this, I much prefer to publish papers with the  $O$ ,  $\Omega$ , and  $\Theta$  notations; I would use other notations like " $(\sim 5n^2)$ " only when faced with a situation that needed it. Why? The main reason is that  $O$ -notation is so universally established and accepted, I would not feel right replacing it by a notation " $(\leq f(n))$ " of my own invention, however logically conceived; the  $O$ -notation has now assumed important mnemonic significance, and we are comfortable with it. For similar reasons, I am not abandoning decimal notation although I find that octal (say) is more logical. And I like the  $\Omega$  and  $\Theta$  notations because they now have mnemonic significance inherited from  $O$ .

Well, I think I have beat this issue to death, knowing of no other arguments pro or con the introduction of  $\Omega$  and  $\Theta$ . On the basis of the issues discussed here, I propose that members of SIGACT, and editors of computer science and mathematics journals, adopt the  $O$ ,  $\Omega$ , and  $\Theta$  notations as defined above, unless a better alternative can be found reasonably soon. Furthermore I propose that the relational notations of Hardy be adopted in those situations where a relational notation is more appropriate.

#### References

- [1] Paul Bachmann, Die Analytische Zahlentheorie. Zahlentheorie, pt. 2 (Leipzig: B. G. Teubner, 1894).
- [2] Paul du Bois-Reymond, "Sur la grandeur relative des infinis des fonctions," Annali di Mat. pura ed applic. (2), 4 (1871), 338-353.
- [3] G. H. Hardy, "Orders of Infinity," Cambridge Tracts in Math. and Math. Physics, 12 (1910; Second edition, 1924).

- [4] G. H. Hardy and J. E. Littlewood, "Some problems of Diophantine approximation," Acta Mathematica 37 (1914), 155-238.
- [5] G. H. Hardy and J. E. Littlewood, "Contributions to the theory of the Riemann zeta function and the theory of the distribution of primes," Acta Mathematica 41 (1918), 119-196.
- [6] Edmund Landau, Handbuch der Lehre von der Verteilung der Primzahlen, 2 vols. (Leipzig: B. G. Teubner, 1909).
- [7] Karl Prachar, Primzahlverteilung (Berlin: Springer, 1957).
- [8] E. C. Titchmarsh, The Theory of the Riemann Zeta-Function (Oxford: Clarendon Press, 1951).
- [9] I. M. Vinogradov, The Method of Trigonometrical Sums in the Theory of Numbers, translated from the 1947 Russian edition by K. F. Roth and Anne Davenport (London: Interscience, no date).

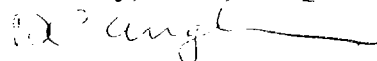
\*\*\*\*\*

11 May 1976

Dear Editor,

The reader of "The Four Russians' Algorithm for Boolean Matrix Multiplication is Optimal in its Class" (News, Vol. 8, No. 1) is advised that its contents are essentially subsumed by "An Algorithm for the Computation of Linear Forms" by J. E. Savage, SIAM J. Comput. Vol. 3 (1974) 150-158, which the author has kindly brought to my attention. Savage presents therein a generalization of the Four Russians' Algorithm, several applications of it, and a counting argument lower bound similar to Moon and Moser's.

Sincerely,



Dana Angluin