

## Chapter 1

# Introduction

### 1.1 Historical notes

The story of the functional equation of associativity begins with Abel. The first paper that he published in Crelle's Journal (volume 1, 1826, pp. 11-15) is entitled "Untersuchung der Function zweier unabhängig veränderlichen Grössen  $x$  und  $y$  wie  $f(x, y)$ , welche die Eigenschaft haben, dass  $f(z, f(x, y))$  eine symmetrische Function von  $x, y$  und  $z$  ist". In this paper, which has been called "the earliest semigroup paper" [Lawson (1996)], Abel showed that if  $f$  is differentiable and satisfies the system of functional equations

$$\begin{aligned} f(x, f(y, z)) &= f(z, f(x, y)) = f(y, f(z, x)) = f(x, f(z, y)) \\ &= f(z, f(y, x)) = f(y, f(x, z)) \end{aligned}$$

then there exists a differentiable and invertible function  $\psi$  such that

$$\psi(f(x, y)) = \psi(x) + \psi(y).$$

Scant attention was given to Abel's paper in the rest of the 19th century. Then, in 1900, Hilbert supplied a strong impetus via the 5th problem of the famous address he delivered at the International Congress of Mathematicians in Paris. (See the English translation which is reprinted in Volume 28 of the Proc. of Symposia in Pure Math., Part I, pp. 1-34.) This problem consists of two parts. In the first part, Hilbert poses the question: "How far Lie's concept of continuous groups of transformations is approachable in our investigations without the assumption of differentiability of the functions". In the second, referring specifically to Abel, he asks: "In how far are the assertions which we can make in the case of differentiable functions true under proper modifications without this assumption?".

Of the two parts of Hilbert's problem, the first is by far the better known and has received much more attention. For the one-dimensional case, solutions were given by L.E.J. Brouwer [1909] and then by E. Cartan [1930]; and the problem as posed was completely solved by A. Gleason, D. Montgomery and L. Zippin in 1952. Work on its ramifications continues to this day. An attack on the second part of Hilbert's problem had to wait for the maturation of the theory of functional equations.

During the first half of the 20th century much of the work on functional equations was sporadic and scattered throughout the mathematical literature. A large number of papers were concerned with the equations that are associated with the names of Cauchy and Jensen; and here and there one finds a paper dealing with the associativity equation

$$T(x, T(y, z)) = T(T(x, y), z). \quad (1.1.1)$$

(For details, see the classic treatise [Aczél (1966)] and the various bibliographies on "Works on Functional Equations" in the journal *Aequationes Mathematicae*.)

The modern theory of the associativity equation (1.1.1) – indeed much of the modern theory of functional equations involving multiplace functions, together with a serious attack on the second part of Hilbert's fifth problem (see [Aczél (1989)]) – begins with the pioneering paper of J. Aczél [1949]. In it he proves the following

**Theorem** *Let  $J$  denote any proper subinterval of the extended real line  $\mathbb{R}$ . Then the function  $A : J \times J \rightarrow J$  is associative, continuous and cancellative if and only if it admits the representation*

$$A(x, y) = a^{-1}(a(x) + a(y)), \quad (1.1.2)$$

where  $a : J \rightarrow \mathbb{R}$  is continuous and strictly monotonic. Moreover,  $J$  must be open or half-open.

Aczél's Theorem improved earlier versions of the representation (1.1.2) by eliminating hypotheses such as commutativity, the presence of an identity, an underlying group structure and, most emphatically, differentiability. It and related results have provided much of the stimulus for the growth of the field of functional equations in the last half-century (see, e.g., the books [Aczél (1966); Kuczma (1968); Dhombres (1979); Targonski (1981); Kuczma (1985); Aczél (1987); Smítal (1988); Aczél and Dhombres (1989);

Kuczma, Choczewski and Ger (1990)], Chapters 5-7 of [Schweizer and Sklar (1983), (2005)], and the journals *Publicationes Mathematicae Debrecen* and *Aequationes Mathematicae*).

While Aczél and his co-workers were developing the theory of functional equations in the spirit of Abel and Cauchy, another group of mathematicians, under the guidance of A.D. Wallace and A.H. Clifford, was extending the work of Brouwer, Cartan et al. from topological groups to topological semigroups (see, e.g., [Faucett (1955); Mostert and Shields (1957); Clifford (1958); Clifford and Preston (1961), (1967); Fuchs (1963); Hofmann and Mostert (1966); Paalman-de Miranda (1970); Carruth, Hildebrandt and Koch (1983)], as well as the recent and interesting surveys [Hofmann (1994); Lawson (1996); Hofmann and Lawson (1996)]).

In 1965, the two roads met when C.-H. Ling [1965], motivated by Aczél's theorem and starting from somewhat different hypotheses, extended the representation (1.1.1) from strictly increasing functions to non-decreasing Archimedean functions (with identity) and then, via the so-called ordinal sums, to all non-decreasing functions. For some extensions of Ling's results see [Krause (1981)] and Section 2.7.

The abovementioned results close one chapter of the study of continuous associative functions on intervals and in a linearly ordered set. The results of Mostert and Shields characterize such functions up to order-isomorphism; and the results of Aczél and Ling yield a universal representation of all such functions. But to say that the story ends here would be analogous to saying that the theory of second order linear differential equations with constant coefficients ends with the general solution of  $ay'' + by' + c = 0$ . For, together with the need for the general solution, there is also a need for particular solutions, particular families of solutions, and solutions having particular properties. Such a need was already evident in the work of T. S. Motzkin [1936], who was interested in finding those associative functions that can be used to define metrics on Cartesian products of metric spaces, and A. Bohnenblust [1940], who was interested in finding the associative functions that characterize norms on certain linear spaces. This need also arose some years later when J. Kampé de Fériet and B. Forte [1967] laid the foundations of the theory of information without probability: here associative functions on the positive half-line, the so-called composition laws, play a crucial role. It is also evident in the studies that eventually led us to write this book.

The origins of this book can be traced back to two papers by the third author and A. Sklar. In the first paper [Schweizer and Sklar (1960)], they laid the foundations of the theory of probabilistic metric spaces, first defined

by K. Menger [1942]. These are generalizations of metric spaces in which the distances between points are described by probability distributions rather than by numbers. (See the book [Schweizer and Sklar (1983), (2005)] as well as the recent survey [Schweizer (2003)]). With each pair of points,  $p$  and  $q$ , in such a space, there is associated a probability distribution function  $F_{pq}$  whose value  $F_{pq}(x)$ , for any real number  $x$ , is usually interpreted as the probability that the “distance” between  $p$  and  $q$  is less than  $x$ . The triangle inequality in such spaces takes the form

$$F_{pr} \geq \tau(F_{pq}, F_{qr}), \quad (1.1.3)$$

where  $\tau$  is a suitable continuous semigroup operation on the space of distribution functions. The most commonly occurring “triangle functions”  $\tau$  have the form

$$\tau_T(F, G)(x) = \sup_{u+v=x} T(F(u), G(v)), \quad (1.1.4)$$

where  $T$  is a continuous “t-norm”, i.e., a suitable continuous semigroup operation on the unit interval  $[0,1]$ . Different triangle functions generally lead to probabilistic metric spaces with different geometric and topological properties; and different t-norms lead to different triangle functions. Thus, in order to penetrate deeply into the structure theory of probabilistic metric spaces, it is necessary to have a repertoire of triangle functions and t-norms at hand. This need was already apparent in the abovementioned first paper and was clearly expressed in the second [Schweizer and Sklar (1961)]. In that paper, Aczél’s Theorem was used to construct several one-parameter families of t-norms, to derive various properties and relations among t-norms, and to analyze the triangle inequality (1.1.3). Last but not least, it was used to relate t-norms to another family of two-place functions, namely the copulas which had been defined earlier by A. Sklar [1959] (see [Sklar (1996a)]). These are functions that link (two-dimensional) probability distribution functions to their one-dimensional margins. They play an important role in the theory of probabilistic metric spaces and, in recent years, have turned out to be increasingly significant in statistics, particularly in the study of concepts and measures of dependence (see [Schweizer (1991)] and the monograph [Nelsen (1999)]).

In 1965, L. Zadeh founded the theory of fuzzy sets (although the original definition goes back to Menger [1951b]) and used the associative functions Maximum and Minimum to define generalized unions and intersections. Subsequently, this work was related to the multivalued logic of

J. Lukasiewicz; and later, in generalizations of the Lukasiewicz theory, as well as in the theory of fuzzy sets itself, Minimum and Maximum were replaced by a  $t$ -norm and its corresponding  $s$ -norm ( $t$ -conorm), respectively [Alsina, Trillas and Valverde (1980), (1983); Dubois (1980); Höhle, private communication; Klement (1980), (1982)]. This work has led to a number of interesting functional equations and inequalities, many of which have been studied by the first author, E. Trillas and their colleagues (see, e.g., [Alsina (1985b), (1988), (1996); Trillas (1980); Trillas and Valverde (1982), (1984); Alsina, Trillas and Valverde (1982)]). The results obtained have found application in the fields of artificial intelligence and cluster analysis [Ruspini (1982); López de Mantaras and Valverde (1984); Zimmermann (1991); Höhle and Klement (1995)] as well as in the theory of synthesis of judgements [Aczél and Saaty (1983); Aczél and Alsina (1984), (1986), (1987); Aczél (1987)]. In addition,  $t$ -norms,  $s$ -norms and other associative functions have come to play a pervasive role in the theory of fuzzy sets, multivalued logics and their ramifications. Listing and discussing all of these matters here would take us too far afield. Fortunately, we do not have to: they are reviewed and developed in detail in the book [Klement, Mesiar and Pap (2000)].

In the last years of the 20th century and in the first years of this century research on  $t$ -norms and related functions has continued at a steady pace. Here we mention the papers [Klement, Mesiar and Pap (2004a), (2004b), (2004c)] in which the salient facts about  $t$ -norms presented in the book [Klement, Mesiar and Pap (2000)], together with some more recent results, are summarized; the lists of open problems [Alsina, Frank and Schweizer (2003); Klement, Mesiar and Pap (2004d)] and lastly, the conference proceedings [Klement and Mesiar (2005)] which give an overview of  $t$ -norms and their applications.

The theory of probabilistic metric spaces has continued to be a source of interesting problems. As indicated previously, in that theory the triangle inequality takes the form (1.1.3), and most triangle functions are induced by semigroup operations on  $[0,1]$ , e.g., as in (1.1.4). In the other direction, many questions concerning such triangle functions can be reduced to questions concerning associative functions on  $[0,1]$ . For example, when  $F$  and  $G$  are step-functions, any functional equation (such as the autodistributivity equation) involving the function  $\tau_T$  defined by (1.1.4) reduces to a functional equation, or a system of functional equations, for  $T$ . Questions of this kind have motivated much of the work of the first author [Alsina (1980), (1981), (1984a)]. Also, in [Frank (1975), (1979), (1991)], the second

author studied a family of binary operations on the space of probability distribution functions which are related to sums of dependent random variables and are induced by copulas. This led him to the functional equation of simultaneous associativity and to the discovery of a one-parameter family of associative copulas which, in recent years, have turned out to be of great importance in certain areas of statistics [Genest and MacKay (1986a), (1986b); Genest (1987); Nelsen (1986), (1991), (1995), (1997)]. Additional examples of this interplay may be found in the text.

One final comment: In the early papers on probabilistic metric spaces, the triangle inequality (1.1.3) had the form  $F_{pr}(x+y) \geq T(F_{pq}(x), F_{qr}(y))$  and, following Menger [1942], the function  $T$  was called a triangular norm, or briefly, a t-norm. Today t-norms arise in many situations where there are no triangles to be found and the name has become an anachronism – and a somewhat misleading one at that. However, this usage of the term is now prevalent in much of the literature, and so, after much agonizing (and after a number of futile attempts to find a reasonable alternative), we have decided not to tamper with it. On the other hand, we have decided to replace the appellation “t-conorm” by “s-norm”.

## 1.2 Preliminaries

In this section we present the notational and other conventions used in this book, as well as some basic prerequisites from the theory of functional equations and inequalities.

Generally, the letters  $u, v, w$  will denote elements of the extended real line  $\mathbb{R} := [-\infty, \infty]$ ;  $x, y, z$  will denote real numbers in the closed unit interval  $I := [0, 1]$ , while  $a, b, c, d$  will be used to denote endpoints of real intervals. In particular, the set of non-negative reals is  $\mathbb{R}^+ := [0, \infty]$ . Real one-place functions will be represented by lower case letters, e.g.,  $f, g, h, t, s$ ; the symbol  $\circ$  will denote composition and  $j_A$  will denote the identity function on the set  $A$  (the reference to  $A$  will be suppressed whenever there is no chance of confusion). When  $\text{Dom } f \subseteq \text{Ran } f$ ,  $f^n$  will denote the  $n$ -th iterate of  $f$ , defined recursively by

$$f^0 = j_{\text{Dom } f} \quad \text{and} \quad f^n = f \circ f^{n-1} \quad \text{for } n \geq 1;$$

and, when it exists, the inverse of  $f$  will be denoted by  $f^{-1}$ .

Binary operations on  $\mathbb{R}$  will be denoted by capital letters, e.g.,  $F, G, H, S, T$ , and will be treated as two-place functions, i.e., we adopt the

functional approach rather than the algebraic one. Also, functions will be allowed to take on the values  $\pm\infty$  at boundary points of their domains.

The most fundamental functional equation is *Cauchy's functional equation*,

$$f(x + y) = f(x) + f(y). \quad (1.2.1)$$

For functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  which satisfy mild regularity conditions (continuity at a single point, or monotonicity, or boundedness from above or below on a set of positive measure, etc.) the general solution of (1.2.1) is

$$f(x) = cx, \quad (1.2.2)$$

where  $c$  is an arbitrary constant. If no regularity condition is assumed, then (1.2.1) admits solutions that are discontinuous everywhere and whose graphs are dense in the plane; indeed, the general solution can be expressed in terms of the values of  $f$  on a Hamel basis of  $\mathbb{R}$ . Moreover, there may be other solutions of Cauchy's equation when (1.2.1) is assumed to hold conditionally, i.e., only for points  $(x, y)$  in a subset of  $\mathbb{R}^2$  [Kuczma (1978), (1985)].

Three other equations of Cauchy type,

$$\begin{aligned} f(x + y) &= f(x)f(y), \\ f(xy) &= f(x) + f(y), \\ f(xy) &= f(x)f(y), \end{aligned} \quad (1.2.3)$$

can be reduced to (1.2.1) by appropriate changes of variables. Under any of the regularity conditions mentioned above, their non-trivial solutions are given by  $f(x) = e^{cx}$ ,  $c \log x$ , and  $x^c$ , on  $\mathbb{R}$ ,  $\mathbb{R}^+$ , and  $\mathbb{R}^+$ , respectively.

Another important functional equation is *Jensen's functional equation*,

$$f\left(\frac{x + y}{2}\right) = \frac{f(x) + f(y)}{2}, \quad (1.2.4)$$

whose general solution, under weak regularity conditions, is

$$f(x) = ax + b,$$

where  $a$  and  $b$  are arbitrary constants.

Equations (1.2.1) and (1.2.4) are closely related to certain basic inequalities. Equation (1.2.1) can be viewed as a special case of either

$$f(x + y) \leq f(x) + f(y) \quad (1.2.5)$$

or

$$f(x + y) \geq f(x) + f(y). \quad (1.2.6)$$

When (1.2.5) or (1.2.6) holds,  $f$  is said to be *subadditive* or *superadditive*, respectively. Equation (1.2.4) leads to the study of

$$f\left(\frac{x + y}{2}\right) \leq \frac{f(x) + f(y)}{2} \quad (1.2.7)$$

and

$$f\left(\frac{x + y}{2}\right) \geq \frac{f(x) + f(y)}{2}. \quad (1.2.8)$$

When (1.2.7) or (1.2.8) holds,  $f$  it is said to be *midpoint* (or *Jensen*) *convex* or *concave*, respectively. Under certain regularity conditions these midpoint properties extend to full convexity and concavity, respectively, i.e., for all  $\lambda$  in  $I$ ,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (1.2.9)$$

or

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y). \quad (1.2.10)$$

There are connections between the above inequalities; for instance, if  $f$  is concave and  $f(0) = 0$ , then  $f$  is subadditive. Most of the important inequalities of classical real analysis are based on one or another of these inequalities.

A functional equation in a single variable which plays a key role in iteration theory is *Schröder's functional equation*,

$$f(g(x)) = sf(x), \quad (1.2.11)$$

where  $g$  is a given function and  $s$  is a constant. Under appropriate conditions on  $g$ , the general solution of (1.2.11) depends on an arbitrary function defined on a subinterval of  $\text{Dom } f$  which can be extended to  $\text{Dom } f$  by way of the iterated equation

$$f(g^n(x)) = s^n f(x).$$

A similar situation prevails in the case of *Abel's functional equation*,

$$f(g(x)) = a + f(x). \quad (1.2.12)$$

For detailed studies of these equations, consult [Aczél (1966), (1987); Aczél and Dhombres (1989); Kuczma (1968), (1985); Dhombres (1979); Kuczma, Choczewski and Ger (1990)]; for the inequalities, see [Hardy, Littlewood and Polya (1952); Beckenbach and Bellman (1965); Marshall and Olkin (1979)].

### 1.3 t-norms and s-norms

The principal objects of our study are the associative binary operations (semigroup operations) on  $I = [0, 1]$  that are order-preserving and commutative and have identity 1 or 0. Throughout, we generally adopt function notation and terminology for algebraic systems. Thus, for instance, a semigroup operation  $T$  on  $I$  is viewed as a two-place function from  $I^2$  to  $I$  that satisfies the functional equation of associativity (1.1.1). This section is devoted to presenting the most basic definitions, terminology, notation, and examples, and to establishing a number of elementary results.

**Definition 1.3.1** A **t-norm** is a two-place function  $T : I^2 \rightarrow I$  (i.e., a binary operation on  $I$ ) which satisfies the following conditions:

(i) On the boundary of  $I^2$ ,

$$T(x, 0) = T(0, x) = 0, \quad (1.3.1a)$$

$$T(x, 1) = T(1, x) = x. \quad (1.3.1b)$$

(ii)  $T$  is non-decreasing in each place, i.e.,

$$T(x_1, y_1) \leq T(x_2, y_2), \text{ whenever } x_1 \leq x_2, y_1 \leq y_2. \quad (1.3.2)$$

(iii)  $T$  is commutative, i.e., for all  $x, y$  in  $I$ ,

$$T(x, y) = T(y, x). \quad (1.3.3)$$

(iv)  $T$  is associative, i.e., for all  $x, y, z$  in  $I$ ,

$$T(T(x, y), z) = T(x, T(y, z)). \quad (1.3.4)$$

Algebraically, a t-norm is a commutative, order-preserving semigroup operation on  $[0, 1]$  with identity 1 and null element 0. Geometrically, the graph of a t-norm is a surface over the unit square which is bounded by the quadrilateral whose vertices are  $(0, 0, 0)$ ,  $(1, 0, 0)$ ,  $(1, 1, 1)$  and  $(0, 1, 0)$ , which rises both horizontally and vertically, and which is symmetric with

respect to the plane  $x = y$ . (See Figure 1.3.1.) Associativity seems to have no simple geometric interpretation.

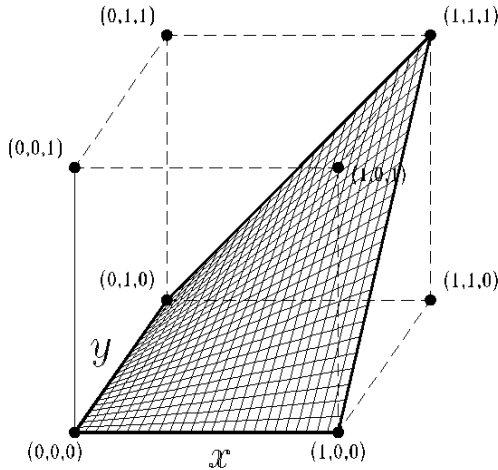


Figure 1.3.1. The graph of a t-norm.

In view of the monotonicity and commutativity of  $T$  and the fact that  $\text{Ran } T \subseteq I$ , the boundary conditions (1.3.1a) and (1.3.1b) can be replaced by the single condition:

(v) For all  $x$  in  $I$ ,

$$T(x, 1) = x.$$

For then  $0 \leq T(x, 0) \leq T(1, 0) = 0$ , so that  $T(x, 0) = 0$ , and the remaining conditions follow from (1.3.3). On the other hand, Examples A.1.1, A.1.2, A.1.3, and A.1.4 of Appendix A show that the conditions (1.3.1), (1.3.2), (1.3.3), and (1.3.4) are independent.

In the presence of continuity the situation changes dramatically. Thus if  $T$  is continuous and satisfies (1.3.1) and (1.3.4), then  $T$  is non-decreasing and commutative, i.e., satisfies (1.3.2) and (1.3.3). (See Lemma 2.1.1 and Corollary 2.1.7.).

**Definition 1.3.2** A t-norm  $T$  is **strict** if it is continuous on  $I^2$  and strictly increasing in each place on  $(0, 1]^2$ , so that

$$T(x_1, y) < T(x_2, y), \quad \text{whenever } x_1 < x_2, y > 0,$$

and

$$T(x, y_1) < T(x, y_2), \quad \text{whenever } x > 0, y_1 < y_2.$$

(1.3.5)

In the sequel we let  $\mathcal{T}$  denote the set of t-norms,  $\mathcal{T}_{Co}$  the set of continuous t-norms, and  $\mathcal{T}_{St}$  the set of strict t-norms. Thus  $\mathcal{T}_{Co}$  consists of the t-norms  $T$  for which  $(I, T)$  is a topological semigroup and, as is readily seen,  $\mathcal{T}_{St}$  those for which the cancellation law also holds. In the literature on topological semigroups, the elements of  $\mathcal{T}_{Co}$  are known as I-semigroups [Paalman-de Miranda (1964); Carruth, et al. (1983)].

The natural ordering on  $I$  induces a partial ordering on the set of all functions from  $I^2$  to  $I$  and, in particular, on the set of all t-norms. Accordingly, we say that  $T_1$  is **weaker** than  $T_2$ , or  $T_2$  is **stronger** than  $T_1$ , and we write  $T_1 < T_2$ , if

$$T_1(x, y) \leq T_2(x, y), \quad \text{for all } (x, y) \text{ in } I^2,$$

and

$$T_1(x_0, y_0) < T_2(x_0, y_0), \quad \text{for some } (x_0, y_0) \text{ in } I^2.$$

If  $T_1 < T_2$  or  $T_1 = T_2$ , we write  $T_1 \leq T_2$ .

The most important t-norms, which we designate by the symbols  $Z$ ,  $W$ ,  $P$ , and  $\text{Min}$ , are given by

$$Z(x, y) = \begin{cases} x, & y = 1, \\ y, & x = 1, \\ 0, & \text{otherwise,} \end{cases}$$

$$W(x, y) = \text{maximum}(x + y - 1, 0), \quad (1.3.6)$$

$$P(x, y) = xy,$$

$$\text{Min}(x, y) = \text{minimum}(x, y).$$

The graphs of these t-norms are depicted in Figure 1.3.2. Note that  $W$ ,  $P$ , and  $\text{Min}$  are in  $\mathcal{T}_{Co}$ , that only  $P$  is in  $\mathcal{T}_{St}$ , and that

$$Z < W < P < \text{Min}. \quad (1.3.7)$$

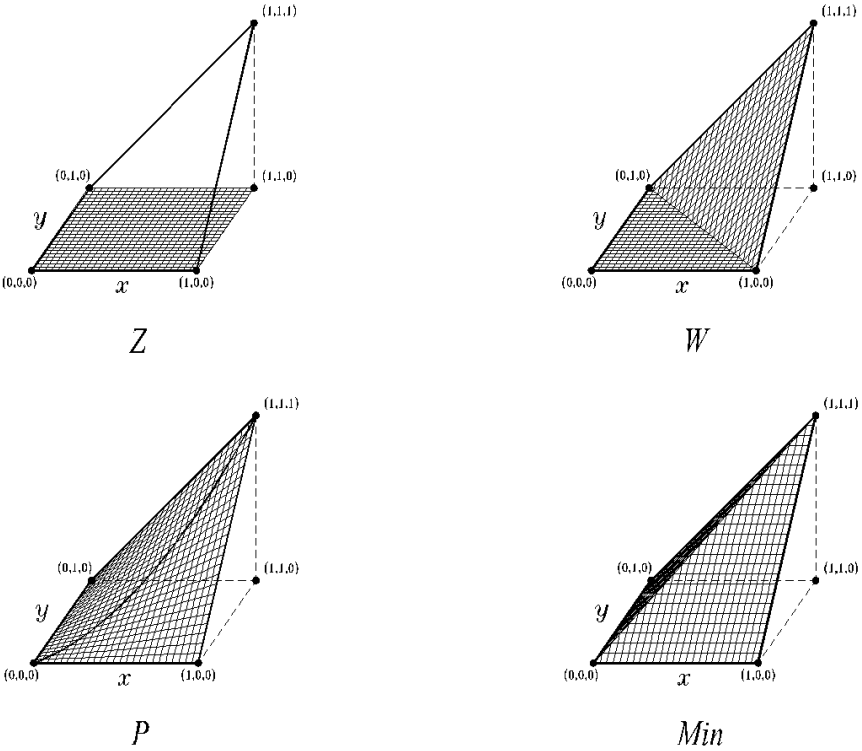


Figure 1.3.2. Graphs of the t-norms  $Z$ ,  $W$ ,  $P$ ,  $Min$ .

**Lemma 1.3.3** *If  $T : I^2 \rightarrow I$  satisfies (1.3.1) and (1.3.2), and in particular if  $T$  is a t-norm, then*

$$Z \leq T \leq Min. \tag{1.3.8}$$

These inequalities follow at once from the inequalities

$$0 \leq T(x, y) \leq T(x, 1) \text{ and } 0 \leq T(x, y) \leq T(1, y),$$

and (1.3.1b). Note, in particular, that  $Z$  is the weakest and  $Min$  the strongest t-norm, or in other words that  $Z$  and  $Min$  are, respectively, the infimum and the supremum of the poset  $(\mathcal{T}, \leq)$ . But since neither the infimum nor the supremum of two associative functions need be associative (see Examples A.9.2, A.9.3 of Appendix A),  $(\mathcal{T}, \leq)$  is not a lattice, nor even a semilattice.

Since  $T$  is associative, for each  $x$  in  $I$ , the **T-powers** of  $x$  may be defined recursively by

$$x^1 = x \text{ and } x^{n+1} = T(x^n, x), \quad (1.3.9)$$

for all positive integers  $n$ . Straightforward inductions yield that, for all  $m, n \geq 1$ ,

$$x^{m+n} = T(x^m, x^n) = T(x^n, x^m), \quad (1.3.10)$$

and

$$x^{mn} = (x^m)^n = (x^n)^m. \quad (1.3.11)$$

Moreover, (1.3.9) can be extended to non-negative integers by defining  $x^0 = 1$  for all  $x \neq 0$ . When this is done, then for all  $x \neq 0$ , (1.3.10) and (1.3.11) extend to all non-negative integers  $m, n$ .

For  $T = \text{Min}$  and  $n \geq 1$ ,  $x^n = x$ ; for  $T = P$ ,  $x^n$  is the ordinary  $n$ -th power of  $x$ ; and for  $T = W$ ,  $x^n = \max(nx - n + 1, 0)$ . (When there is a possibility of confusion, we write  $x_T^n$  rather than merely  $x^n$ .)

For any binary operation  $B$  on  $I$ , let  $B^* : I^2 \rightarrow I$  be the function defined by

$$B^*(x, y) = 1 - B(1 - x, 1 - y). \quad (1.3.12)$$

It is evident that  $B^*$  is non-decreasing in each place, commutative, or associative if and only if  $B$  has the corresponding property, and that  $B^{**} = B$ . Moreover the graphs of  $B$  and  $B^*$  are reflections of each other in the point  $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ .

For any t-norm  $T$  we have that  $T^*(x, 0) = x$  and  $T^*(x, 1) = 1$ . Thus the association (1.3.12) yields a dual class of semigroup operations on  $I$ , having identity 0 and null element 1. This motivates

**Definition 1.3.4** An **s-norm** is a two-place function  $S : I^2 \rightarrow I$  which satisfies the monotonicity, commutativity, and associativity conditions (1.3.2), (1.3.3), (1.3.4) and the boundary conditions

$$S(x, 0) = S(0, x) = x, \quad S(x, 1) = S(1, x) = 1. \quad (1.3.13)$$

An s-norm is **strict** if it is continuous on  $I^2$  and strictly increasing in each place on  $[0, 1]^2$ .

We let  $\mathcal{S}, \mathcal{S}_{Co}$ , and  $\mathcal{S}_{St}$  denote the set of s-norms, continuous s-norms and strict s-norms, respectively.

The s-norms associated with the t-norms listed in (1.3.6) are

$$\begin{aligned} \text{Min}^*(x, y) &= \text{Max}(x, y), \\ P^*(x, y) &= x + y - xy, \\ W^*(x, y) &= \text{Min}(x + y, 1), \\ Z^*(x, y) &= \begin{cases} x, y = 0, \\ y, x = 0, \\ 1, \text{ otherwise.} \end{cases} \end{aligned} \tag{1.3.14}$$

The graphs of these s-norms are depicted in Figure 1.3.3.

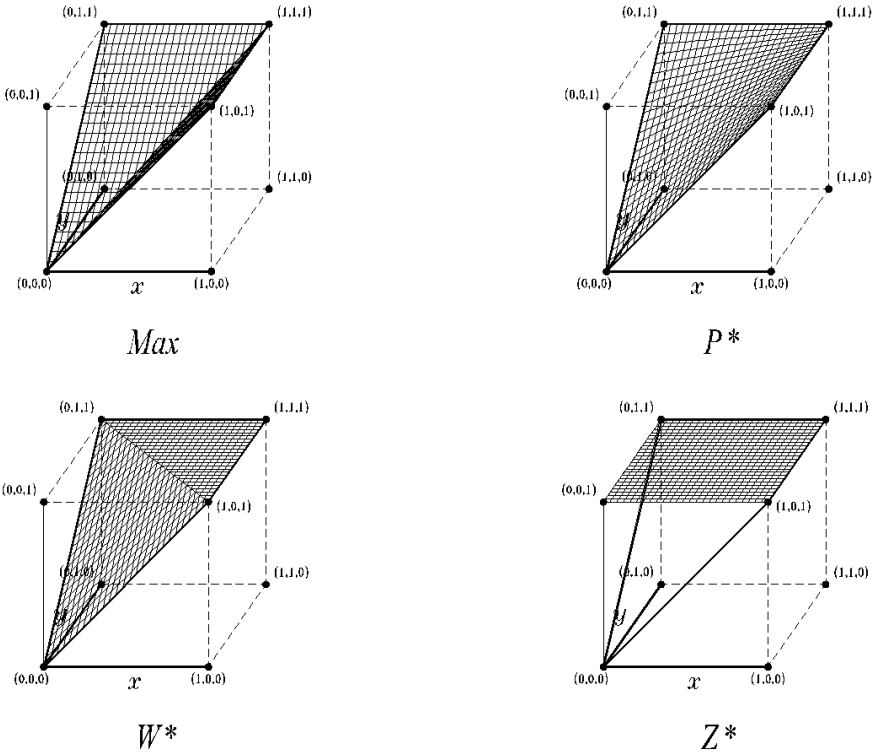


Figure 1.3.3. Graphs of the s-norms  $\text{Max}, P^*, W^*, Z^*$ .

From the previous discussion it is immediate that  $T$  is a t-norm if and only if  $T^*$  is an s-norm. It therefore follows that for any statement concerning t-norms there is a corresponding dual statement for s-norms. For instance,  $T$  is in  $\mathcal{T}_{Co}$  (resp., in  $\mathcal{T}_{St}$ ) if and only if  $T^*$  is in  $\mathcal{S}_{Co}$  (resp.,  $\mathcal{S}_{St}$ );  $T_1 < T_2$  iff  $T_1^* > T_2^*$ ; the s-norms displayed in (1.3.14) satisfy

$$\text{Max} < P^* < W^* < Z^*; \quad (1.3.15)$$

and, for any s-norm  $S$ ,

$$\text{Max} \leq S \leq Z^*. \quad (1.3.16)$$

Note also that, since  $\text{Min} < \text{Max}$ ,

$$T < S, \quad (1.3.17)$$

for any t-norm  $T$  and any s-norm  $S$ .

The restrictions of associative functions to the line  $y = x$  play a fundamental role in the sequel. Accordingly, the **diagonal** of a t-norm  $T$  is the function  $\delta_T : I \rightarrow I$  defined by

$$\delta_T(x) = T(x, x). \quad (1.3.18)$$

The iterates of  $\delta_T$  are the functions  $\delta_T^n$  defined recursively by

$$\delta_T^0 = j_I \quad \text{and} \quad \delta_T^{n+1} = \delta_T \circ \delta_T^n.$$

**Lemma 1.3.5** *The diagonal  $\delta_T$  of the t-norm  $T$  has the following properties:*

- (a)  $\delta_T$  is non-decreasing, with  $\delta_T(0) = 0$  and  $\delta_T(1) = 1$ ;
- (b)  $\delta_T(x) \leq x$ , for all  $x$  in  $I$ ;
- (c) for any  $x$  in  $(0,1)$ , the sequence  $\{\delta_T^n(x)\}$  is non-increasing and  $0 \leq \lim_{n \rightarrow \infty} \delta_T^n(x) < 1$ ;
- (d) in terms of the  $T$ -powers defined in (1.3.9),

$$\delta_T^n(x) = x^{2^n}. \quad (1.3.20)$$

*Proof.* Properties (a) and (b) are obvious; (c) now follows from the inequalities  $\delta_T^{n+1}(x) = \delta_T(\delta_T^n(x)) \leq \delta_T^n(x)$  and  $0 \leq \delta_T^n(x) \leq x$ ; and a simple induction yields (d).  $\square$

An **idempotent** of a t-norm  $T$  is an element  $x$  of  $I$  for which  $T(x, x) = x$ , i.e.,  $\delta_T(x) = x$ . The elements 0 and 1 are idempotents of every t-norm.

Note that if  $x$  is an idempotent of  $T$ , then for every  $y \geq x$ ,  $x = T(x, x) \leq T(x, y) \leq T(x, 1) = x$ , i.e.,  $T(x, y) = T(y, x) = x$ , which immediately yields

**Lemma 1.3.6** *If  $\delta_T(x) = x$  for all  $x$  in  $I$ , then  $T = \text{Min}$ .*

At the other extreme, the t-norms for which 0 and 1 are the only idempotents are of special importance. Clearly, for any such  $T$ ,

$$\delta_T(x) < x, \quad \text{for } 0 < x < 1. \quad (1.3.21)$$

In particular, every  $T$  in  $\mathcal{T}_{St}$  satisfies (1.3.21) since, in this case,  $\delta_T(x) = T(x, x) < T(x, 1) = x$ .

**Definition 1.3.7** A t-norm  $T$  is **Archimedean** if, for any  $x, y$  in  $(0,1)$ , there exists a positive integer  $m$  such that  $x^m < y$ .

Combining Lemma 1.3.5 with the fact that the sequence  $\{x^m\}$  is non-increasing immediately yields

**Lemma 1.3.8** *A t-norm  $T$  is Archimedean if and only if*

$$\lim_{n \rightarrow \infty} \delta_T^n(x) = 0, \quad \text{for } 0 < x < 1. \quad (1.3.22)$$

We let  $\mathcal{T}_{Ar}$  denote the set of Archimedean elements of  $\mathcal{T}_{Co}$ .

**Theorem 1.3.9** *A continuous t-norm is Archimedean if and only if it does not have any interior idempotents, i.e., if and only if it satisfies (1.3.21). In particular, every strict t-norm is Archimedean, i.e.,  $\mathcal{T}_{St} \subset \mathcal{T}_{Ar}$ .*

*Proof.* It is sufficient to show that (1.3.21) and (1.3.22) are equivalent when the diagonal  $\delta$  is continuous. First, if  $\delta(x) = x$  for some  $x$  in  $(0,1)$ , then  $\delta^n(x) = x$  for all  $n$ , contradicting (1.3.22). To prove the converse, observe that  $\delta(\lim_{n \rightarrow \infty} \delta^n(x)) = \lim_{n \rightarrow \infty} \delta(\delta^n(x)) = \lim_{n \rightarrow \infty} \delta^n(x)$  for all  $x$ , whence (1.3.21) and Lemma 1.3.5(c) together yield (1.3.22).  $\square$

For discontinuous t-norms, the Archimedean property is stronger than (1.3.21); moreover, even strict monotonicity of a t-norm does not imply the Archimedean property. A full discussion of these facts, including the extension of Theorem 1.3.9 to all t-norms, is presented in Section 2.5. Note that  $Z$  and  $W$  are Archimedean but not strictly increasing.

The preceding statements concerning diagonals of t-norms and the Archimedean property have their counterparts for s-norms. These all follow readily from the fact that, for any s-norm  $S$ ,

$$\delta_S(x) = 1 - \delta_{S^*}(1 - x). \quad (1.3.23)$$

In particular, for s-norms the inequalities (1.3.19) and (1.3.21) are reversed, the sequence  $\delta_S^n$  is non-decreasing, the Archimedean inequality is  $x^m > y$  (where the  $x^m$  are  $S$ -powers), and the limit in (1.3.22) is 1. Theorem 1.3.9 is valid verbatim for s-norms, upon replacing the inequality in (1.3.21) by  $\delta_S(x) > x$ .

The **sections** of a t-norm  $T$  are the functions obtained by fixing  $T$  in one place. Thus, for each  $y$  in  $I$ , the  $y$ -**section**  $\sigma_y$  is the function defined on  $I$  by

$$\sigma_y(x) = T(x, y). \quad (1.3.24)$$

In view of the commutativity, the  $x$ -sections of  $T$  are the same family of functions. Each  $y$ -section is non-decreasing, with  $\sigma_y(0) = 0$  and  $\sigma_y(1) = y$ .

## 1.4 Copulas

A class of binary operations, which are related to t-norms and which are of great importance in probability and statistics, are the copulas introduced by A. Sklar [1959].

**Definition 1.4.1** A (two-dimensional) **copula** is a two-place function  $C : I^2 \rightarrow I$  which satisfies the boundary conditions (1.3.1) and the monotonicity condition

$$C(x_1, y_1) - C(x_2, y_1) - C(x_1, y_2) + C(x_2, y_2) \geq 0, \quad (1.4.1)$$

whenever  $x_1 \leq x_2$ ,  $y_1 \leq y_2$ .

It is easy to check that  $W$ ,  $P$  and  $\text{Min}$  satisfy (1.4.1) and are thus copulas. Several basic properties of copulas are collected in

**Lemma 1.4.2** *If  $C$  is a copula, then*

$$0 \leq C(x_2, y_2) - C(x_1, y_1) \leq x_2 - x_1 + y_2 - y_1, \quad (1.4.2)$$

whenever  $x_1 \leq x_2$ ,  $y_1 \leq y_2$ . Thus every copula is non-decreasing in each place and Lipschitz continuous. Moreover, for any copula  $C$ ,

$$W \leq C \leq \text{Min}. \quad (1.4.3)$$

*Proof.* To obtain the first inequality in (1.4.2), let  $x_1 = 0$  in (1.4.1); let  $y_1 = 0$  in (1.4.1) and then relabel  $y_2$  as  $y_1$ ; add the resulting inequalities. Similarly, to obtain the second, let  $x_2 = 1$  in (1.4.1), let  $y_2 = 1$  in (1.4.1), then relabel  $y_1$  as  $y_2$  and add these inequalities. Finally, letting  $x_2 = y_2 = 1$  in (1.4.1), and combining the result with Lemma 1.3.3 yields (1.4.3).  $\square$

In the statistical literature, the largest and smallest copulas, Min and  $W$ , in (1.4.3) are generally referred to as the **Fréchet-Hoeffding bounds** [Fréchet (1935), (1951); Hoeffding (1940)].

Copulas were so-named because they link bivariate probability distributions to their margins. The exact connection is given by the following basic result, first announced in [Sklar (1959)]; a proof is given in [Schweizer and Sklar (1974)] (see also [Sklar (1996a); Nelsen (1999)]).

**Theorem 1.4.3** *Let  $H$  be a bivariate probability distribution with margins  $F$  and  $G$ . Then there is a copula  $C$  such that*

$$H(u, v) = C(F(u), G(v)), \quad (1.4.4)$$

for all  $u, v$  in  $\mathbb{R}$ . If  $F$  and  $G$  are continuous,  $C$  is unique; otherwise,  $C$  is uniquely determined on  $(\text{Ran } F) \times (\text{Ran } G)$ . In the other direction, for any univariate distributions  $F, G$  and any copula  $C$ , the function  $H$  defined by (1.4.4) is a bivariate distribution with margins  $F$  and  $G$ .

Copulas therefore provide a natural setting for the study of properties of distributions with fixed margins. Note, in particular, that a copula is itself a continuous bivariate distribution on  $I^2$  with uniform margins. The extension of the copula notion and Theorem 1.4.3 to higher dimensions, also due to A. Sklar [1959], will be presented in Section 4.4.

In the next theorem, we collect the key results which establish the role of copulas in the theory of probability and in mathematical statistics. For the sake of brevity, we restrict our attention to continuous distributions.

Let  $X$  and  $Y$  be random variables, defined on a common probability space and taking values in  $\mathbb{R}$ , with continuous distributions  $F_X, F_Y$  and joint distribution  $H_{XY}$ , i.e.,

$$F_X(u) = \text{Pr}(X \leq u), \quad F_Y(v) = \text{Pr}(Y \leq v),$$

$$H_{XY}(u, v) = Pr(X \leq u, Y \leq v);$$

and denote by  $C_{XY}$  the (unique) copula guaranteed by Theorem 1.4.3, so that

$$H_{XY}(u, v) = C_{XY}(F_X(u), F_Y(v)). \quad (1.4.5)$$

**Theorem 1.4.4** *Let  $X, Y, F_X, F_Y, H_{XY}$ , and  $C_{XY}$  be as in the preceding paragraph. Then:*

- (a)  $X$  and  $Y$  are independent if and only if  $C_{XY} = P$ .
- (b)  $Y = f(X)$  a.s., where  $f$  is strictly increasing a.s. on  $\text{Ran } X$ , if and only if  $C_{XY} = \text{Min}$ .
- (c)  $Y = f(X)$  a.s., where  $f$  is strictly decreasing a.s. on  $\text{Ran } X$ , if and only if  $C_{XY} = W$ .
- (d) If  $f$  and  $g$  are strictly increasing a.s. on  $\text{Ran } X$  and  $\text{Ran } Y$ , respectively, then

$$C_{f(X), g(Y)} = C_{XY}.$$

- (e) If  $f$  and  $g$  are strictly decreasing a.s. on  $\text{Ran } X$  and  $\text{Ran } Y$ , respectively, then

$$C_{f(X), Y}(x, y) = y - C_{XY}(1 - x, y), \quad (1.4.6a)$$

$$C_{X, g(Y)}(x, y) = x - C_{XY}(x, 1 - y), \quad (1.4.6b)$$

$$C_{f(X), g(Y)}(x, y) = x + y - 1 + C_{XY}(1 - x, 1 - y). \quad (1.4.7)$$

- (f)  $C_{XY}$  is the restriction to  $I^2$  of the joint distribution of the probability transforms  $F_X(X)$  and  $F_Y(Y)$ .
- (g) For continuous distributions  $F$  and  $G$ , let  $X_1 = F^{-1}(F_X(X))$  and  $Y_1 = G^{-1}(F_Y(Y))$ . Then  $F_{X_1} = F$ ,  $F_{Y_1} = G$ , and  $C_{X_1 Y_1} = C_{XY}$ .

Parts (a)-(c) are classical results of M. Fréchet [1951, 1957, 1958]. The proofs of the remaining parts are straightforward. These two theorems show that much of the study of joint distributions can be reduced to the study of copulas. In particular, (1.4.5) and part (d) of Theorem 1.4.4 imply that it is precisely the copula that captures those properties of the joint distribution which are invariant under a.s. strictly increasing transformations. Thus, for example, the quantity

$$\sigma(X, Y) = 12 \int_0^1 \int_0^1 |C_{XY}(x, y) - xy| dx dy$$

is a measure of monotone dependence of the random variables  $X$  and  $Y$  [Schweizer and Wolff (1981)]. Moreover, many well-established stochastic notions are equivalent to simple properties of copulas. For instance, the notion of stochastic dominance translates to pointwise order: If  $H_1 = H_{X_1Y_1}$  and  $H_2 = H_{X_2Y_2}$  are bivariate distributions with equal margins, i.e.,  $F_{X_1} = F_{X_2}$  and  $F_{Y_1} = F_{Y_2}$ , then  $(X_2, Y_2)$  is said to dominate  $(X_1, Y_1)$  if  $H_2 \geq H_1$ , which is the case precisely when their copulas satisfy  $C_2 \geq C_1$ .

In the years between the appearance of the seminal paper by A. Sklar [1959] and the papers by the third author and E.F. Wolff [1976, 1981], most of the results concerning copulas were obtained in connection with problems arising from the study of probabilistic metric spaces. Then, in the decade 1980-1990, research on copulas and their applications grew markedly and culminated in an international conference which was held in Rome in 1990. Since then, four additional international conferences have been held – in Seattle in 1993, in Prague in 1996, in Barcelona in 2000 and in Quebec City in 2004; and since the turn of the century there has been an explosion of copula-centered activity, fueled largely by the significant role that these functions play in the area of finance and risk management. This monograph is not the place to review the extensive and rapidly expanding literature. Instead, we refer the reader to the proceedings of the above-mentioned conferences [Dall’Aglia, Kotz and Salinetti (1991); Rüschemdorf, Schweizer and Taylor (1996); Beneš and Štěpán (1997); Cuadras, Fortiana and Rodriguez-Lallena (2002); Genest et al. (2005)] for details; to the paper [Schweizer (1991)] for a survey of developments from 1959 to 1989; to the papers [Dall’Aglia (1991)] and [Sklar (1996a)] for interesting comments on the early history of the subject; to the monograph [Nelsen (1999)] for a comprehensive and eminently readable introduction to the subject, together with an extensive bibliography; and to the books [Hutchinson and Lai (1990); Joe (1997); Drouot Mari and Kotz (2001); Cherubini, Luciano and Vecchiato (2004)].

In this monograph we are primarily concerned with associative copulas. A t-norm that satisfies (1.4.1) is a copula, and in view of Lemma 1.4.2 and the paragraph immediately preceding Definition 1.3.2, an associative copula is a t-norm. Many of the important copulas and families of copulas are associative. However, there are commutative copulas that are not associative, and hence not t-norms; and there are t-norms that satisfy (1.4.3) but not (1.4.1), and hence are not copulas. (See Examples A.2.1 and A.2.2 of Appendix A.)

The following characterization of associative copulas is due to R. Moynihan [1978] (see also [Schweizer and Sklar (1983), (2005), Theorem 6.3.2]).

**Theorem 1.4.5** *A t-norm  $T$  is a copula if and only if it satisfies the Lipschitz condition*

$$T(x_2, y) - T(x_1, y) \leq x_2 - x_1, \quad \text{whenever } x_1 \leq x_2. \quad (1.4.8)$$

*Proof.* It is obvious from (1.4.2) that any copula satisfies (1.4.8). In the other direction, if  $T$  satisfies (1.4.8), then  $T$  is continuous. Choose  $x_1 \leq x_2$  and  $y_1 \leq y_2$ . Since  $T(0, y_2) = 0 \leq y_1 \leq y_2 = T(1, y_2)$ , there exists a number  $z$  in  $I$  such that  $T(z, y_2) = y_1$ . Hence, by (1.3.3), (1.3.4), and (1.4.8),

$$\begin{aligned} T(x_2, y_1) - T(x_1, y_1) &= T(x_2, T(z, y_2)) - T(x_1, T(z, y_2)) \\ &= T(T(x_2, y_2), z) - T(T(x_1, y_2), z) \\ &\leq T(x_2, y_2) - T(x_1, y_2), \end{aligned}$$

which is (1.4.1). □

The **dual copula** of the copula  $C$  is the function  $C^\wedge$  defined on  $I^2$  by

$$C^\wedge(x, y) = x + y - C(x, y). \quad (1.4.9)$$

Dual copulas are motivated by a natural probabilistic interpretation: in the notation of (1.4.5),

$$\begin{aligned} Pr(X \leq u \text{ or } Y \leq v) &= F_X(u) + F_Y(v) - C_{XY}(F_X(u), F_Y(v)) \\ &= C^\wedge_{XY}(F_X(u), F_Y(v)). \end{aligned} \quad (1.4.10)$$

The properties of  $C^\wedge$  listed below follow readily from Lemma 1.4.2.

**Lemma 1.4.6** *If  $C$  is a copula and  $C^\wedge$  is its dual copula, then:*

- (a)  $\text{Ran } C^\wedge = I$ , i.e.,  $C^\wedge$  is a binary operation on  $I$ ;
- (b)  $C^\wedge$  satisfies the boundary conditions (1.3.13);
- (c)  $C^\wedge$  is non-decreasing in each place;
- (d)  $C^\wedge$  is commutative if and only if  $C$  is commutative.

It is emphatically not true in general that  $C^\wedge$  is an s-norm when  $C$  is a t-norm. This is so because  $C^\wedge$  usually fails to be associative. The canonical copulas  $W$ ,  $P$ , and  $\text{Min}$ , however, are exceptions: for each of

these,  $C^\wedge = C^*$ . The complete answer to the question as to when  $C$  and  $C^\wedge$  are simultaneously associative was given by the second author in [Frank (1979)] where he showed that this is the case if and only if  $C$  belongs to the important family of copulas that now bears his name. The details are given in Section 3.1.

In [Alsina, Nelsen and Schweizer (1993)] the notion of a copula was extended to that of a **quasi-copula**. The original definition was rather unwieldy. Then, in [Genest, Quesada-Molina, Rodriguez-Lallena and Sempi (1999)] it was proved that  $Q : I^2 \rightarrow I$  is a quasi-copula if and only if  $Q$  satisfies (1.3.1), (1.3.2) and (1.4.8). hence, in view of Theorem 1.4.5, every associative quasi-copula is a copula.

With the discovery of the new definition, the pace of research on quasi-copulas quickened. A brief review of recent results is given in [Nelsen (2005)]. Here we only note that the set of 2-quasi-copulas is the Dedekind-MacNeille completion of the poset of 2-copulas [Nelsen and Úbeda-Flores, to appear].

Finally, we note that the set of copulas is convex.