Convolutions of Gamma Distributions

Frank Massey

The convolution of two functions *f*(*t*) and *g*(*t*) defined for positive *t* is (*f* \* *g*)(*t*) = . The density function of the sum *S* + *T* of two independent, non-negative random variables *S* and *T* is the convolution of the density functions of *S* and *T*. A gamma random variable is a non-negative random variable with density function *γ*(*t*;*α*,*λ*) = *λt*‑1*e*-*λt*/*Γ*(**) for *t* > 0 for some positive numbers *α* and *λ*. A sum of independent gamma random variables has density function *A*(*t*;*******,****λ***) = *γ*(*t*;*α*1,*λ*1) \* *γ*(*t*;*α*2,*λ*2) \* … \* *γ*(*t*;*αn*,*λn*) for some vectors ****** = (**1,...,*n*) and ***λ*** = (**1,...,*n*) of positive numbers. We look at ways to find properties of *A*(*t*;*******,****λ***) despite the fact that it is usually not possible to express *A*(*t*;*******,****λ***) in terms of elementary functions.

**1. Basic Properties of Convolutions.**

If *f*(*t*) and *g*(*t*) are two functions that are defined for *t* > 0 and integrable over any finite interval, then their convolution (*f* \* *g*)(*t*) is defined by

(1) (*f* \* *g*)(*t*) = *f*(*s*)*g*(*t-s*)*ds* = *f*(*t-r*)*g*(*r*)*dr*

We write *f*(*t*) \* *g*(*t*) or *f* \* *g* instead of (*f* \* *g*)(*t*) if it is more convenient. Convolution has the following properties

*f* \* *g* = *g* \* *f* commutative

*f* \* (*g* \* *h*) = (*f* \* *g*) \* *h* associative

*f* \* (*g* + *h*) = (*f* \* *g*) + (*f* \* *h*) convolution distributes over addition

(2) L(*f* \* *g*) = L(*f*) L(*g*) Laplace transform changes convolution to a product

(3) [*eηtf*1(*t*)] \* \* [*eηtfn*(*t*)] = *eηt* [*f*1(*t*) \* \* *fn*(*t*)]

 Multiplication by an exponential distributes over convolution

Here L(*f*) is the Laplace transform of *f*(*t*).

(4) L(*f*) = L(*f*)(*s*) = ∞

Commutativity and distributivity follow easily from (1). Associativity follows from the Laplace transform formula (2) or from applying the definition and interchanging the order in the resulting double integral. (3) also follows from the Laplace transform formula and the fact that the Laplace transform of an exponential *eηt* times a function is the transform of the function shifted by *η*. Another way to prove (3) is to note that for *n* = 2 it follows easily from the definition. General *n* is proved by induction. To prove (2) one uses (4) and (1) and interchanges the order of integration in the resulting double integral as follows.

 L(*f* \* *g*)(*s*) = ∞*e*-*st*(*f* \* *g*)(*t*) *dt* = ∞*e*-*stf*(*r*)*g*(*t-r*) *drdt*

 = ∞∞*e*-*stf*(*r*)*g*(*t-r*) *dtdr* = ∞∞*e*-*stf*(*r*)*e*-*sug*(*u*) *dudr* = L(*f*) L(*g*)

**Example 1.**

 1 \* = *dr* =

**Example 2.** =

where 1 is convoluted with itself *n* times. This is proved using Example 1 and induction on *n*.

**Example 3.**

 \* = \*

= =

**Example 4.**

(5) *αΓα* \* *βΓβ* = *αβΓαβ*

where *α* and *β* are positive real numbers and

 *Γ*(**) = ∞*tα*-1*e*-*t* *dt*

is the gamma function. If *α* is a positive integer, then *Γ*(**) = (*α* - 1)!, so the gamma function is the factorial function translated one unit to the right and extended to real values. The formula *n*! = *n*(*n*-1)! extends to the gamma function, i.e. *Γ*(*α* + 1) = *αΓ*(*α*). For half-integers, the gamma function can also be expressed in terms of factorials, i.e. *Γ* = *π* and if *n* is an integer then *Γ*(*n* + ) = …*π* .

To prove (5) one uses (2) to get

(6) L*αΓαβΓβ* = L*αΓα* L*βΓβ*

For L*αΓα* one has

 L*αΓα* = ∞*e*-*st* *αΓα* *dt* = ∞*e*-*r* *αΓαα* *dt* = *α*

So (6) becomes

 L*αΓαβΓβ* = *α* *β* = *αβ* = L*αβΓαβ*

which proves (2).

**Example 5.** Equation (5) extends to

(7) *αΓα* \* *αΓα* \* … \* *αΓα* = *Γ*

for a vector of positive numbers ****** = (**1,...,*n*). Here

(8) |******| = *α*1 + ... + *αn*

is the sum of the *αj*.

**Example 6.** If we multiply (7) by *e*-*λt* and distribute this over the convolution using (3) we get

(9) *αλΓα* \* *αλΓα* \* … \* *αλΓα* = *λΓ*

which we shall use in a moment when we get to sums of gamma random variables.

**2. Applications of Convolutions.**

Convolutions have a number of applications. They arise in the solution of linear constant coefficient differential equations. For example, the solution to the initial value problem

 + *kx* = *f*(*t*) *x* = 0 when *t* = 0

can be written as

 *x*(*t*) = *e*-*kt* \* *f*(*t*)

 More generally, the solution to the initial value problem

 + *an*-1 + … + *a*1 + *a*0*x* = *f*(*t*)

 *x*(*n*-1)(0) = *x*(*n*-2)(0) = … = *x*'(0) = *x*(0) = 0

can be written as

 *x*(*t*) = *φ*(*t*) \* *f*(*t*)

where *φ*(*t*) is the solution to

 + *an*-1 + … + *a*1 + *a*0*x* = 0

 *x*(*n*-1)(0) = 1 *x*(*n*-2)(0) = … = *x*'(0) = *x*(0) = 0

In fact,

 *φ*(*t*) = L-1

where L-1 is the inverse Laplace transform.

Convolutions also arise in probability. Let *S* and *T* be independent, non-negative, continuous random variables with density functions *f*(*t*) and *g*(*t*) respectively which are zero for *t* < 0. Then the density function of *S* + *T* is (*f* \* *g*)(*t*). To see this, let *H*(*t*) = *Pr*{*S* + *T*  *t*} be the cumulative distribution function of *S* + *T*. One has

 *H*(*t*) = *Pr*{*S* + *T*  *t*} = *Pr*{(*S*,*T*)  *A*} = *φ*(*r*,*s*) *drds*

where *φ*(*r*,*s*) is the joint distribution function of *S* and *T* and *A* = {(*r*,*s*): *r* ≥ 0, *s* ≥ 0, *r* + *s*  *t*}. Since *S* and *T* are independent one has *φ*(*r*,*s*) = *f*(*r*)*g*(*s*). So

 *H*(*t*) = *f*(*r*)*g*(*s*) *drds* = *F*(*t*-*s*)*g*(*s*) *ds*

where *F*(*s*) is the cumulative of *S*. So the density function of *S* + *T* is

 *H*'(*t*) = = *F*(0)*g*(*t*) + *∂∂* = = (*f* \* *g*)(*t*)

Note *F*(0) = 0 since *f*(*t*) = 0 for *t* < 0.

 More generally, let *T*1, …, *Tn* be independent, non-negative, continuous random variables with density functions *f*1(*t*), …, *fn*(*t*) respectively which are zero for *t* < 0. Then the density function of *S* = *T*1 + … + *Tn* is

 *A*(*t*) = *f*1(*t*) \* … \* *fn*(*t*)

**3. Sums of Gamma Random Variables.**

The purpose of these notes is to look at properties of the density function of the sum of independent gamma random variables. In probability, a random variable *T* has a *gamma distribution* (or is a *gamma random variable*) if its density function *fT*(*t*) has the form

(10) *fT*(*t*) = *γ*(*t*;*α*,*λ*) = *λ*‑*λΓ*

for some positive numbers *α* and *λ*. *α* is sometimes called the shape parameter and *λ* the scale parameter. *α* = 1 is an important special case. In that case

 *fT*(*t*) = *γ*(*t*;1,*λ*) = *λλ*

and *T* is called an *exponential random variable*. Exponential random variables are used to model times for events to occur. For example, the times between customer arrivals at a bank or service times for customers at a bank teller or the time for a radioactive nucleus to decay.

 Gamma random variables have a number of applications. Like the special case of an exponential random variable, they are used to model times for events to occur in situations where an exponential random variable doesn't seem to adequately describe the situation.

Another important application of gamma random variables is to describe the variance when a random sample is selected from a population with a normal distribution. If *X* is normally distributed with mean 0 and variance *σ*2, then *X*2 has a gamma distribution with *α* = ½ and *λ* = 1/(2*σ*2), i.e.

 *fX*(*x*) = *πσe*-*x*2/(2*σ*2) ⇒ *fX*2(*s*) = *γ*(*t*;,*σ*) = *σσπ*

More generally, if *X*1, …, *Xn* are independent normal random variables each with mean 0 and variance *σ*2, then (*X*1)2 + … + (*Xn*)2 has a gamma distribution with *α* = *n*/2 and *λ* = 1/(2*σ*2), i.e.

 *fXj*(*x*) = *πσe*-*x*2/(2*σ*2) ⇒ *fX*12+…+*Xn*2(*s*) = *γ*(*t*;,*σ*) = *σσΓ*

Finally, if *X*1, …, *Xn* are independent normal random variables each with mean *μ* and variance *σ*2, then if we let = (*X*1 + … + *Xn*)/*n*, then (*X*1 - )2 + … + (*Xn* - )2 has a gamma distribution with *α* = (*n*-1)/2 and *λ* = 1/(2*σ*2).

The special case of a gamma distribution where *α* = *n*/2 where *n* is a positive integer and *λ* = ½ is called a *χ*2 distribution with *n* degrees of freedom, i.e.

 *fχ*2(*n*)(*s*) = *γ*(*t*;,) = *Γ*

In particular, (*X*1 ‑ )2/*σ*2 has a *χ*2 distribution with *n* – 1 degrees of freedom.

 Let *S* = *T*1 + … + *Tn* be the sum of independent gamma random variables *T*1, …, *Tn* where *Tj* has parameters *j* and *λj*. Let ****** = (**1,...,*n*) and ***λ*** = (**1,...,*n*) be the vectors of shape and scale parameters of the *Tj*. The density function of *S* is

(11) *A*(*t*) = *A*(*t*;*******,****λ***) = *λααλΓα* \* *λααλΓα* \* … \* *λααλΓα*

 Sums of independent gamma random variables are used in similar circumstances to gamma random variables. They are used to model the time for a series of events to occur where the time for each event separately to occur is modeled by a gamma random variable. They are also used to model sums of squares analogous to variances in situations where we are selecting objects randomly from populations with different means and variances and the populations we are selecting them from are also selected at random.

One can not express *A*(*t*) in terms of elementary functions except if the *αj* are all integers or if the *λj* are all equal. The purpose of these notes is look at ways to determine properties of *A*(*t*).

**4. Properties of the Density Function.**

 To begin, let's look at the properties of a single gamma random variable, as opposed to a sum of gamma random variable. In other words, what does the graph of *f*(*t*;*α*,*λ*) given by (10) look like. Let's first look at the case where *λ* = 1, i.e.

 *fT*(*t*) = *γ*(*t*;*α*,1) = ‑*Γ*

Note that for *t* > 0 one has *fT*'(*t*) = (*α* - 1 – *t*)*tα*-2*e*-*t*/*Γ*(*α*). If 0 < *α* < 1 then *f*(*t*) is decreasing for *t* > 0 and *fT*(*t*) → ∞ as *t* ↓ 0 and *fT*(*t*) ↓ 0 as *t* → ∞. If *α* = 1 then *fT*(*t*) = *e*-*t*. If *α* > 1 then *f*(0) = 0 and as *t* increases, *fT*(*t*) increases to a maximum at *t* = *α* - 1 and then decreases back to 0 as *t* → ∞. At the right are the graphs of *y* = *t*-1/2*e*-*t*/*π*, *y* = *e*-*t* and *y* = *t*2*e*-*t*.

 In going from *γ*(*t*;*α*,1) to get *γ*(*t*;*α*,*λ*) we stretch or shrink the graph of *γ*(*t*;*α*,1) horizontally by a factor of 1/*λ* and stretch or shrink vertically by a factor of *λ*. So the graph of *γ*(*t*;*α*,*λ*) is similar to the graph of *γ*(*t*;*α*,1).

 Now let's turn to the properties of *A*(*t*;*******,****λ***). We can get an indication of what to expect by looking at the special case where all the *λj* are equal. In that case *A*(*t*;*******,****λ***) is just the density function of a single gamma random variable with shape parameter equal to the sum of the *αj*.

**Proposition 1.** Let ****** = (*α*1,...,*αn*) be a vectors of positive numbers and ***λ*** = (*λ*,...,*λ*) be a vector of *n* identical real numbers. Then

(12) *A*(*t*;******, (*λ*,...,*λ*)) = *λλΓ*

**Proof.** This is just the equation (9) multiplied by *λ*|******|.

So, in the case where all the *λj* are equal, then the sum of the *αj* controls the behavior of *A*(*t*;*******,****λ***). We shall show that to a certain extent, this is true in the general case.

**5. Small *t*.**

We start with upper and lower bounds for *A*(*t*;*******,****λ***).

**Proposition 2.** Let *μ*1 = *min*{**1,...,*n*} and *μn* = *max*{**1,...,*n*}. Then

(13) ***λα***‑*μΓ****α*** ≤ *A*(*t*;******,***λ***) ≤ ***λα***‑*μΓ****α***

where

(14) ***λ*** = (*λ*1)**1(*λ*2)**2…(*λn*)*n*

**Proof.** Since *e*‑*μnt* ≤ *e*‑*λkt* ≤ *e*‑*μ*1*t* for all *k* we have

***λ****μ****α*** *μααμΓα* \* *μααμΓα* \* … \* *μααμΓα*

≤ *λααλΓα* \* *λααλΓα* \* … \* *λααλΓα*

≤ ***λ****μ****α*** *μααμΓα* \* *μααμΓα* \* … \* *μααμΓα*

In other words,

(15) ***λ****μ****α*** *A*(*t*;******,(*μn*,...,*μn*)) ≤ *A*(*t*;******,***λ***) ≤ ***λ****μ****α*** *A*(*t*;******,(*μ*1,...,*μ*1))

Using Proposition 1 one has

 *A*(*t*;******,(*μn*,...,*μn*)) = *μμΓ*

and similarly for *A*(*t*;******,(*μ*1,...,*μ*1)). So (15) becomes

 ***λ****μ****α*** *μμΓ* ≤ *A*(*t*;******,***λ***) ≤ ***λ****μ****α*** *μμΓ*

and (13) follows.

Since *e*-*μnt* ≈ 1 and *e*-*μ*1*t* ≈ 1 for small *t* it follows from Proposition 2 that

*A*(*t*;******,***λ***) ≈ ***λα****Γ****α***

for small *t*, i.e. *A*(*t*;******,***λ***) is approximately proportional to *t*|***α***|-1 for small *t*.

 A better small *t* approximation to *A*(*t*;******,***λ***) is

*A*(*t*;******,***λ***) ≈ *λλΓ*

where

 *λ* = *λλ*

This has a relative error of O(*t*2) for small *t*.

 Another consequence of Proposition 2 is that *A*(*t*;******,***λ***) goes to zero at a rate no bigger than a constant times *t*|******|-1*e*-*μ*1*t* and no smaller than a constant times *t*|******|-1*e*-*μnt* as *t* → ∞.

**6. Unimodality.**

 The next issue that we are interested in is the increasing/decreasing behavior of *A*(*t*) = *A*(*t*;******,***λ***). So far we have been able to show the following.

**Proposition 3.**

(16) If 0 < |*α*| < 1 then *A*(*t*) is decreasing for *t* > 0 and *A*(*t*) → ∞ as *t* ↓ 0 and *A*(*t*) ↓ 0 as *t* → ∞.

(17) If |*α*| = 1 then *A*(*t*) is decreasing for *t* > 0 and *A*(*t*) → ***λ****Γ****α*** as *t* ↓ 0 and *A*(*t*) ↓ 0 as *t* → ∞.

(18) If |*α*| > 1 and ∈*αj* ≤ 1 where *S* = {*j*: *αj* < 1}, then *A*(0) = 0 and as *t* increases, *A*(*t*) increases to a maximum and then decreases back to 0 as *t* → ∞.

**Remark.** We would like to show that (19) holds true without the restriction ∈*αj* ≤ 1 where *S* = {*j*: *αj* < 1}.

The proof of (17) and (18) uses the following representation of *A*(*t*;******,***λ***).

**Proposition 4.** Let *u*0 = 1, *un* = 0, *Γ*(******) = *Γ*(**1)*Γ*(**2) … *Γ*(*n*),*du* = *du*1…*dun*-1. Then

(19) *A*(*t*;******,***λ***) = ***λ****t*|**|-1*H*(*t*;******,******)

where

(20) *H*(*t*;******,******) = *Ωh*(*u*;******) *e*-*t*(*λ*#*u*) *du*

(21) *h*(*u*;******) = *Γ*(******)-1 (*u*0 - *u*1)**1-1(*u*1 - *u*2)**2-1…(*un*-1-*un*)*n*-1

(22) *λ*#*u* = *λ*1(*u*0-*u*1) + *λ*2(*u*1-*u*2) + *λ*3(*u*2-*u*3) + … + *λn*‑1(*un*-2-*un*-1) + *λn*(*un*-1-*un*)

(23) *Ω* = {(*u*1,…,*un*‑1): 0 ≤ *u*1 ≤ 1, 0 ≤ *u*2 ≤ *u*1,…, 0 ≤ *un*-1 ≤ *un*-2}

**Remark.** In the case *n* = 2 the formulas (20) – (23) become

(24) *H*(*t*;******,******) = (1 - *u*1)**1-1(*u*1)**2-1 *e*-*t*(*λ*#*u*) *du*

**Proof.** (19) is a special case of a general representation of convolutions of *n* functions, namely

(25) *f*1(*t*) \* \* *fn*(*t*) = *tn*-1 *Ω*

To prove (25) in the case *n* = 2 we start out with the definition of convolution *f*1(*t*) \* *f*2(*t*) = *‑* and make the change of variables *s* = *tu* which gives *f*1(*t*) \* *f*2(*t*) = *t*  which is (25) for *n* = 2. The case of general *n* follows by induction. Applying (25) with *fj*(*t*) = [*Γ*(*j*)]-1*tj*-1*e*-*λjt* gives

 *A*(*t*;******,***λ***) = ***λ****tn*-1*Γ*(******)-1 *Ω*((*t*(*u*0 - *u*1))**1-1*e*-*λ*1*t*(*u*0 - *u*1)(*t*(*u*1 - *u*2))**2-1*e*-*λ*2*t*(*u*1 - *u*2)…(*t*(*un*-1-*un*))*n*-1*e*-*λnt*(*un*-1 - *un*) *du*

The proposition follows from this.

**Proof of (16) and (17).** By Proposition 4 one has *A*(*t*;******,***λ***) = ***λ****t*|**|-1*H*(*t*;******,******) where *H*(*t*;******,******) = *Ωλ*. If |***α***| ≤ 1 then *t*|**|-1 is decreasing in *t*. Also, *e*-*t*(*λ*#*u*) is decresing in *t*. Since *h*(*u*;******) ≥ 0 it follows that *H*(*t*;******,******) is decreasing in *t*. So *A*(*t*;******,***λ***) is decreasing in *t*.

The proof of (18) involves some general results on unimodal functions.

**Definition 1.** Let *f*(*t*) be a function that is integrable over any finite interval of the real line and let *F*(*t*) = . Then *τ* is a *mode* for *f*(*t*) if *F*(*t*) is convex on - ∞ < *t* ≤ *τ* and *F*(*t*) is concave on *τ* ≤ *t* < ∞. *f*(*t*) is *unimodal* at *τ* if *τ* is a mode for *f*(*t*). *f*(*t*) is *unimodal* if *f*(*t*) has mode.

**Remarks.** (1) If *f*(*t*) is unimodal at *τ*, then the left and right derivatives of *F*(*t*) exist everywhere except possibly at *τ*. If we redefine *f*(*t*) to be equal to some value between the left and right derivatives of *F*(*t*) for those values of *t* for which it is not between these values, then *f*(*t*) is non‑decreasing for - ∞ < *t* < *τ* and non-increasing for *τ* < *t* < ∞. Conversely, if *f*(*t*) is non‑decreasing for - ∞ < *t* < *τ* and non-increasing for *τ* < *t* < ∞ then *f*(*t*) is unimodal at *τ*.

(2) A mode of a unimodal function need not be unique. However, the modes of a unimodal function form a closed interval.

**Example 7.** The gamma density function *λ*‑*λΓ* is unimodal for all positive *α* and *λ*.

**Example 8.** (16) and (17) show that *A*(*t*;******,***λ***) is unimodal if |***α***| ≤ 1.

In some cases where one can not show directly that a non-negative function is unimodal one can show that it is unimodal either by showing that is log-concave or it is the convolution of a unimodal function with a log-concave function.

**Definition 2.** A function *y* = *f*(*t*) is *concave* if it assumes values in the range - ∞ ≤ *y* < ∞ and *f*(*αt* + (1-*α*)*s*) ≤ *αf*(*t*) + (1-*α*)*f*(*s*) for all *s* and *t* and 0 ≤ *α* ≤ 1. For the purposes of this we use the defintions (- ∞) + *y* = *y* + (- ∞) = - ∞ and (- ∞)*y* = *y*(- ∞) = - ∞ if - ∞ ≤ *y* < ∞.

**Definition 3.** A function *f*(*t*) is *log-concave* if *f*(*t*) ≥ 0 for all *t* and log[*f*(*t*)] is concave. For the purposes of this definition we define log[0] = - ∞.

**Remarks.** It is not hard to see that *f*(*t*) is log-concave if and only if *f*(*t*) ≥ 0 for all *t* and *f*(*αt* + (1‑*α*)*s*) ≤ *f*(*t*)*αf*(*s*)1‑*α* for all *s* and *t* and 0 ≤ *α* ≤ 1. A log-concave function is unimodal.

**Example 9.** The gamma density function *λ*‑*λΓ* is log-concave if *α* ≥ 1.

A very important connection between log-concavity and unimodality are the following two Theorems whose proof can be found in S. Dharmadhikari and J.D. Kumar book *Unimodality, Convexity and Applications* published by Academic Press, 1988.

**Theorem 5.** The convolution of log‑concave functions is again log-concave.

**Theorem 6 (Ibramov).** The convolution of a log-concave function with a unimodal function is again unimodal.

 So *A*(*t*;******,***λ***) is log-concave if *αj* ≥ 1 for all *j* and unimodal if *αj* < 1 for at most one *j*.

**Corollary 5.** Suppose one If *α*1 + … + *αk* ≤ 1 then *A*(*t*;******,***λ***) is unimodal.

**Proof of (18).** Since convolution is commutative and associative, we may reorders the (*αj*, *λj*) so that the *αj* increase with *j*. Suppose *k* is such that *αj* < 1 for *j* ≤ *k* and *αj* ≥ 1 for *j* > *k*. By the hypothesis of (18) one has *α*1 + … + *αk* ≤ 1. By (16) and (17) one has *A*(*t*;(*α*1, …, *αk*),(*λ*1, …, *λk*)) decreasing. Also by Example 9 and Theorem 5 *A*(*t*;(*αk*+1, …, *αn*),(*λk*+1, …, *λn*)) is log-concave. So *A*(*t*;******,***λ***) = *A*(*t*;(*α*1, …, *αk*),(*λ*1, …, *λk*)) \* *A*(*t*;(*αk*+1, …, *αn*),(*λk*+1, …, *λn*)) is unimodal by Theorem 6.

 The case where we haven't proven *A*(*t*;******,***λ***) is unimodal is where |***a***| > 1 but *αj* < 1 for all *j*. For example, *A*(*t*;(½,¾),(1,2)) = *Γ* \* *Γ*.