How averaging individual curves transforms their shape

Mathematical analyses with application to learning and forgetting curves

Murre, J.M.J.

DOI
10.1016/j.jmp.2023.102816

Publication date
2023

Document Version
Final published version

Published in
Journal of Mathematical Psychology

Journal of Mathematical Psychology 117 (2023) 102816

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

How averaging individual curves transforms their shape: Mathematical analyses with application to learning and forgetting curves

Jaap M.J. Murre

Department of Psychology, University of Amsterdam, Netherlands

A R T I C L E I N F O

Article history:
Received 29 June 2022
Received in revised form 15 August 2023
Accepted 1 September 2023
Available online 18 September 2023

Keywords:
Averaging
Learning curve
Forgetting curve
Power law
Power function
Arithmetic averaging
Geometric averaging

A B S T R A C T

This paper demonstrates how averaging over individual learning and forgetting curves gives rise to transformed averaged curves. In an earlier paper (Murre and Chessa, 2011), we already showed that averaging over exponential functions tends to give a power function. The present paper expands on the analyses with exponential functions. Also, it is shown that averaging over power functions tends to give a log power function. Moreover, a general proof is given how averaging over logarithmic functions retains that shape in a specific manner. The analyses assume that the learning rate has a specific statistical distribution, such as a beta, gamma, uniform, or half-normal distribution. Shifting these distributions to the right, so that there are no low learning rates (censoring), is analyzed as well and some general results are given. Finally, geometric averaging is analyzed, and its limits are discussed in remedying averaging artefacts.

1. Introduction

As has been noted in the psychological literature since the 1950s (Bakan, 1954; Sidman, 1952), arithmetic averaging (i.e., when we take the average as usual) over individual learning or forgetting curves from different subjects is *not* a neutral operation. It often results in a mathematical transformation of the underlying individual curves (Murre & Chessa, 2011), which was first established through simulations (Anderson & Tweney, 1997; Brown & Heathcote, 2003). The averaged curve is a valid mathematical description of the individual curves only if they have specific mathematical shapes (Estes, 1956; Myung, Kim, & Pitt, 2000) or if the to-be-averaged curves are highly similar.

This problem takes center stage in the debate surrounding the shape of the learning curve, where a 'Power Law of Learning' has been proposed and defended by some (Anderson & Schooler, 1991; Newell & Rosenbloom, 1981; Wixted & Ebbesen, 1991), whereas others have argued that individual curves are not distributed according to a power function but rather an exponential function (Heathcote, Brown, & Mewhort, 2000) and have adduced evidence that averaging over these causes an artefactual power function (Anderson & Tweney, 1997). We have proven mathematically that this is indeed a possibility: averaging over individual exponential functions may result in a power function, either exactly or in the limit (Murre & Chessa, 2011).

The same debate continues in the forgetting literature and started with the seminal work of Hermann Ebbinghaus, who was the first researcher to fit an equation to forgetting data, namely in his habilitation thesis of 1880. On the basis of a number of well-argued assumptions, he constructed a differential equation, solving it by deriving a forgetting equation. Curious is that despite his excellent analysis, he deems the result merely of 'empirical significance' (German: "nur ganz empirische Bedeutung", p.63) without an important theoretical meaning. This is probably also signaled by the fact that he put the whole section 12 (p. 57–63), which describes this analysis, in square brackets. He is clearly not very attached to his own equation either, because in the eventual publication of his thesis as a book in 1885, he has replaced it with a different one (Ebbinghaus, 1913/1885; see Murre & Dros, 2015, for an in depth discussion).

Hermann Ebbinghaus used himself as a subject and hence the problem of averaging over subjects did not occur in his work. He also went to great lengths to find stimuli that were equally difficult to learn and to forget. Unfortunately, because he learned rows of nonsense syllables in a fixed order, the middle syllables were more difficult to remember, still causing an uneven distribution in the items learned (Murre & Dros, 2015). This makes it unlikely that there was no averaging artefact in his forgetting curve; averaging over trials with items of different levels of learning could still have transformed the shape.

Sidman (Iversen, 2021; 1952) was one of the first researchers to argue that despite the possible theoretical implications of the mathematical shape of certain averaged curves, great care should

E-mail address: jaap@murre.com.

https://doi.org/10.1016/j.jmp.2023.102816

be taken when making inferences from them without further information. A power function shape of an averaged learning curve, for example, might be due to subjects all having exactly the same power curve but also because they have exponential learning curves with individual learning rates that follow a gamma distribution (Murre & Chessa, 2011). The two resulting averaged curves might be indistinguishable both in practice and in theory, even though their origins are quite different and would lead one to propose diverging underlying mechanisms.

Whether one does or does not want to draw inferences from the shape of averaged learning or forgetting curves, it is crucial to be aware of averaging artefacts, especially if the original data is not available. The raw data would allow one to examine and fit the individual curves, which in all cases is to be preferred. This paper expands our earlier mathematical analyses (Murre & Chessa, 2011), which focused on averaging exponential learning curves with a learning rate that follows the gamma, half-normal, or uniform distribution. Here, I will also consider the beta distribution, which is of considerable interest because of its great flexibility in shape, being able to mimic other distributions (e.g., uniform, left and right triangular, [bimodal] arcsine distributions, and the Bernoulli and normal distributions in the limit). I will also extend the analyses to averaging over individual power functions and logarithmic functions. The power function has often been proposed as a learning or forgetting function (e.g., Anderson & Schooler, 1991; Ebbinghaus, 1880; Newell & Rosenbloom, 1981; Wickelgren, 1974). The logarithmic function is of interest and has been proposed by some researchers as a forgetting function (e.g., Ebbinghaus, 1885/1964, replaces the power function by a logarithmic function). I will, furthermore, study what happens if the distribution of learning (or forgetting) rates is shifted to the right, so that there are no slow learners (censoring). An analysis within our framework of the pros and cons of using the geometric mean instead of the usual arithmetic one completes the analyses. The results are discussed in general terms in the next section, with most details of the mathematical derivations and proofs given in the Methods section, which has been placed in Appendix A.

2. Mathematical analyses: Results

2.1. Arithmetic averaging

Three types of individual learning functions will be considered, which I will refer to as *base functions*: exponential, power, and logarithmic. Each of these has a learning rate parameter, determining how quickly learners progress. The analyses here apply equally here to forgetting and any other response curves that can be modeled by the base functions; for brevity, I will only mention learning. When considering learning, one should bear in mind that many factors influence its progress. Here, we will only focus on the role of learning time and ignore all other factors. For a recent in-depth discussion of factors influencing the learning process see, for example, Chechile (2018, Chapter 11). When averaging over subjects, it is likely that not all subjects have exactly the same learning rate. We will, therefore, assume that the variation from low to high rates may be captured by some statistical distribution. Performance on a typical learning task can be expressed either as percentage correct, which would be an increasing function of time (assuming regularly spaced learning trials), or as percentage incorrect.

For the learning rates we will use three well-known distributions: the gamma distribution (which has the exponential distribution as a special case), the half-normal distribution, and the beta distribution (which has the uniform distribution as a special case). As is illustrated in Fig. 1, both the gamma and the beta distribution are quite flexible in terms of possible shapes. The beta distribution is suitable for cases where on a priori grounds the learning rate must remain below some upper bound (which is 1 by default but can be changed through a reparameterization). None of the distributions allow negative values.

The results of the analyses are given in Table 1 with the details of their derivation relegated to Appendix A. In some cases, the results in Table 1 represent a limiting form: as time t (in suitable time units) increases, the shape of the averaged function will approach this form. The exact forms are given in Table A.1 in the Methods section in Appendix A. The limiting forms are shown here to emphasize the similarity in the equations for the averaged curves, which are as follows.

(1) *Averaging over exponential functions gives power functions*. This includes our earlier findings (Murre & Chessa, 2011), which are here extended with the beta distribution. The fact that the beta distribution gives such an elegant shape as a limiting form is perhaps surprising, because it includes a bimodal shape for certain parameter values (see Fig. 1). It should be remarked, however, that the convergence for the bimodal shapes can indeed be rather slow. As was already shown in our earlier paper, the result for the uniform distribution shows a rather fast convergence (e.g., Murre & Chessa, 2011, Figure 2).

(2) *Averaging over power functions gives logarithmic power functions*. Again, we observe the same general shape for all averaged curves, despite the differences in the distributions. We must also note that the limiting form of the beta distributions also exhibits a rather slow convergence.

(3) *Averaging over logarithmic functions retains their shape*. This result was already noted by Estes (1956). Note that we use the unconditioned form, $c - \log(x\,t)$, as a base function and keep the equation shape similar to the other two, even though the actual curve is now descending. It would be much better to use a well-condition version of the logarithm here such as $(\log(tx+1)+1)^{-1}$, which value would stay between 0 and 1, but I could only find a closed-form solution in the limit for the arithmetic average with the uniform distribution, yielding a log power log form (see the proof in Appendix A). This is consistent with averaging over power functions yielding a log power shape, because we have wrapped a logarithmic function inside a power function with exponent -1.

As shown in Appendix A, an interesting general result with the unconditioned logarithmic base function can be obtained, namely that arithmetic averaging over base functions of the shape $c - \log(x\,t)$ always gives an averaged curve of shape:

$$\begin{aligned} P_A(t) &= \int_0^\infty f(x)\,(c - \log(xt))\,dx \\ &= c - \log(t) - \log\left(GM_f\right) \end{aligned}$$

where GM_f is the geometric mean of distribution f. For example, the geometric mean of the beta distribution is $e^{\psi^{(0)}(a)-\psi^{(0)}(a+b)}$ (Vogel, 2020), which is why the constant $-B$ in Table 1 is $\psi^{(0)}(a) - \psi^{(0)}(a+b)$, where $\psi^{(0)}(z)$ is the digamma function.

3. Discussion

In Table 1, we can easily spot that are consistent patterns in the equations of the averaged curves, which for the unconditioned logarithmic function could be proven mathematically; I was not able to find a general proof that accounts for the observed similarities in shape for the exponential and power base functions. The results are illustrated in Fig. 2, which visualizes how averaging affects the different base functions.

As can be seen in Fig. 2a, there are quite a few curves where learning progresses very slowly. These contribute to a slowed-down descent in the averaged curve, which in case of exponential

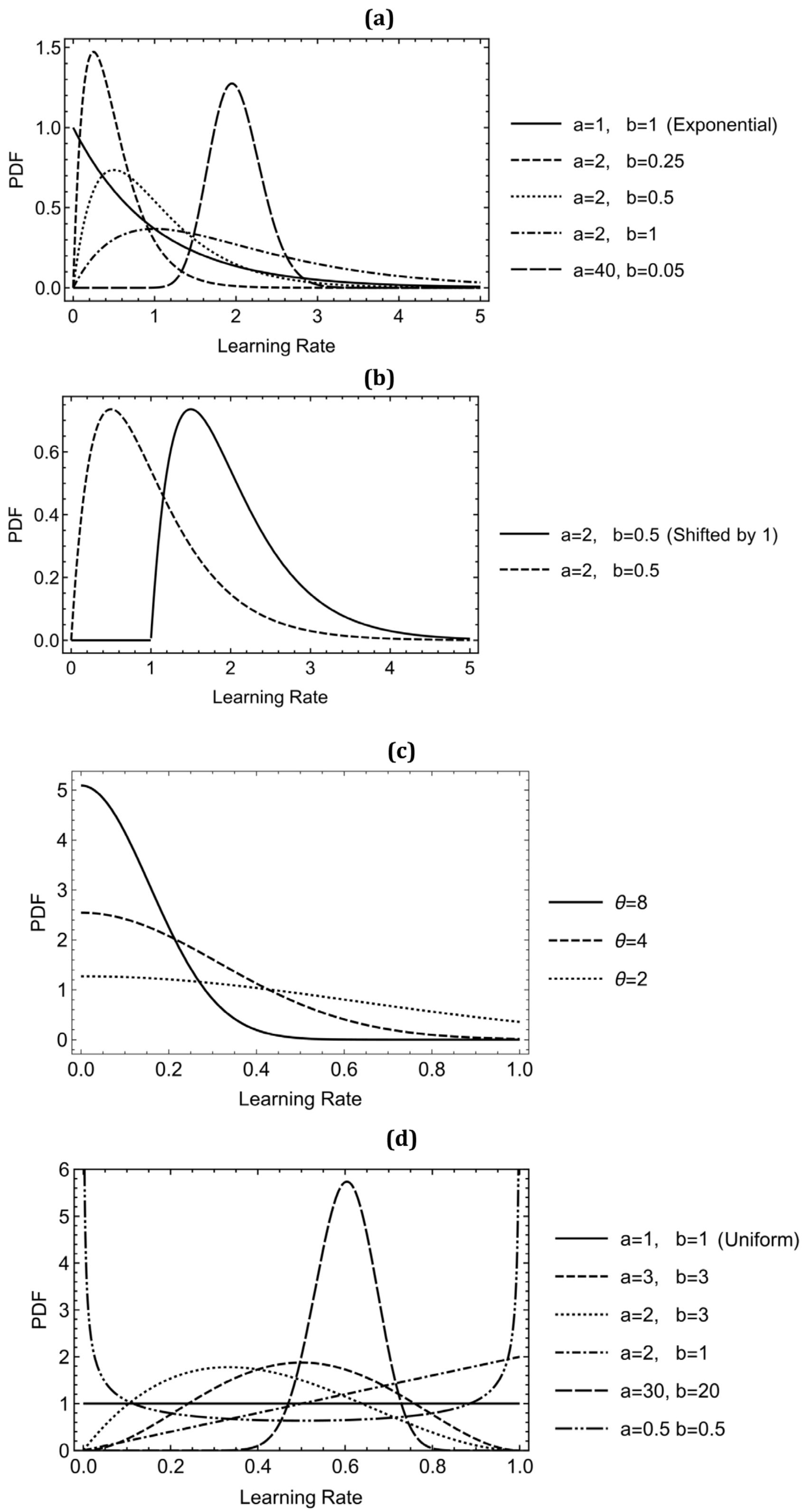

Fig. 1. Illustration of the three learning rate distributions used in the paper. (a) Gamma distribution, $f(x) = \frac{b^{-a} x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a)}$, with $x > 0$. (b) Gamma distribution with a variant shifted by 1. (c) Half-normal distribution, $f(x) = \frac{2\theta e^{-\frac{\theta^2 x^2}{\pi}}}{\pi}$, with $x > 0$. (d) Beta distribution, $f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, with $0 < x < 1$.

base functions is a power function. The reason for the many slow learners is that distributions often have a relatively high density close to 0. This is clearly the case with the exponential and uniform distribution. What would happen if we shifted the distribution to higher learning rates, thereby removing the very slow learning rates? The effects of this can be analyzed, among others, by shifting distributions to the right. This is illustrated for the gamma distribution in Fig. 1b.

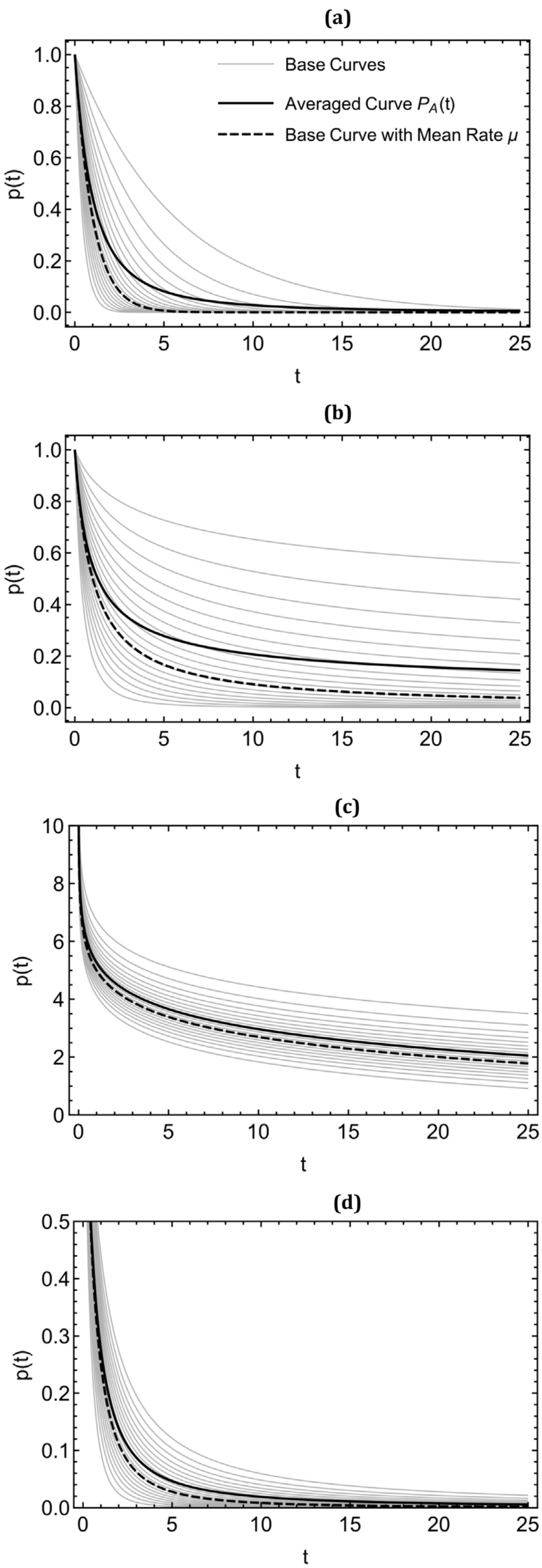

Fig. 2. Illustration of the averaging effects with a mixing gamma distribution with parameters $a = 2$ and $b = 0.5$. The base curves (grey solid line) are spaced regularly with rate values taken in steps of 5% quantiles, starting with the 5% quantile (low learning rate) and ending with the 95% quantile (high learning rate). For the three base functions, the averaged curve from Table 1, $P_A(t)$, is plotted (black solid line) with the base function with the mean rate $\mu = 1$ of the gamma distribution (black dashed line). (a) Exponential base functions. (b) Power base functions. (c) Logarithm base functions. (d) Power base function with the gamma distribution $f(x - 1)$ shifted with $z = 1$.

Appendix A gives the derivation of the following general result: If we shift the learning rate distribution to the right with a distance z, we obtain a rate distribution $f(x - z)$. Mixing such a distribution with exponential base functions gives $P_z(t) = e^{-zt} P_A(t)$, where $P_z(t)$ is the averaged function using the shifted distribution and $P_A(t)$ is the averaged function using the unshifted distribution (i.e., the equations shown in Table A.1 in the Appendix A). Similarly, for power functions we have $P_z(t) = (t+1)^{-z} P_A(t)$. The effects of shifting the gamma distribution with $z = 1$ is shown in Fig. 2.d for power base functions. I was not able to derive useful equations for the logarithmic base function with a shifted distribution.

Some authors have suggested using geometric averaging rather than (ordinary) arithmetic averaging to retain the shape of exponential and power functions (Anderson & Tweney, 1997; Estes, 1956). The sample geometric mean is:

$$GM = \sqrt[n]{x_1 x_2 \ldots x_n} = \exp\left(\sum_{i=1}^{n} \frac{\log x_i}{n}\right), \text{ where } x > 0$$

For a mixing distribution $f(x)$ with (arithmetic) mean μ, assuming a domain of 0 to ∞ (or the applicable domain of $f(x)$), we calculate the averaged curve $P_A(t)$ as follows (Vogel, 2020), using $g(x)$ as the base function:

$$P_A(t) = \exp\left(\int_0^{\infty} f(x)\, log\,(g(x))\, dx\right)$$

As is shown in Appendix A, for exponential base functions, this gives exactly $P_A(t) = e^{-\mu t}$, and for power base functions $P_A(t) = (t+1)^{-\mu}$. In other words, for exponential and power base functions, no matter what the learning rate distribution is, the geometric average will be a curve that has as the mean of the distribution the learning rate parameter. In many cases, this will be the desired result, which is why one should consider geometric above arithmetic averaging in case the individual curves are exponential or power functions.

The result above relies on certain properties of exponential and power functions, which the logarithmic function does not possess. For logarithmic base functions, I have not been able to derive useful results for either the gamma, half-normal, or beta distribution. In case of logarithmic base functions, it makes most sense to use arithmetic averaging, which at least retains the logarithmic shape (see above). Another case in which geometric averaging does not give useful closed-form solution, is when base functions are of the shape $1 - g(x)$, or more generally, $A + Bg(x)$. This gives expressions of the shape $\log(A + Bg(x))$ in the integral, which are difficult to integrate to closed-form solutions. In Appendix A, I give the geometrically averaged curve of $1 - e^{-xt}$ with a uniform rate distribution. A close-form solution exists, but it does not preserve the exponential shape, indicating that indeed geometric averaging may be of limited use in these cases. This confirms mathematically what had already been demonstrated through simulations by Brown and Heathcote (2003), namely that the original shape of exponential functions is not necessarily exactly preserved with geometric averaging in case of a non-zero asymptote in the base function. Though the theoretical limits of geometric averaging are worth investigating, the reader should bear in mind that in case the base function is expected to be of the form $y = 1 - e^{-xt}$, an easy solution to still obtain undistorted geometric averaging results is to work with $1 - y$ rather then y. Then, calculate the sample geometric mean M by summing all observations $log(1-y)$ and take the exponential, as in the formula above. We can then use $1 - M$ as the final result.

In conclusion, I agree with Ebbinghaus (1880) that fitted curves are first of all of empirical significance. As such, they can be very useful and important in practice, for example, to predict

Table 1
Functions – including some limiting forms for high t – resulting from arithmetic averaging over the base functions with different mixing distributions on rate parameter x.

Mixing distributions	Base functions ($\boldsymbol{t}$ is time, $\boldsymbol{x}$ is the learning rate)		
	$1-e^{-xt}$	$1-(1+t)^{-x}$	$c-\log(xt)$
Gamma	$1-(1+bt)^{-a}$	$1-(1+b\log(1+t))^{-a}$	$c-\log(bt)+A$
Exponential	$1-(1+bt)^{-1}$	$1-(1+b\log(1+t))^{-1}$	$c-\log(bt)+\gamma$
Half-normal	$1-\frac{2\theta}{\pi}t^{-1}$ [a]	$1-\frac{2\theta}{\pi}\log^{-1}(t)$ [a]	$c-\log\left(\frac{\sqrt{\pi}}{2}\frac{1}{\theta}t\right)+\frac{\gamma}{2}$
Beta	$1-Gt^{-a}$ [b]	$1-G\log^{-a}(t+1)$ [b]	$c-\log(t)+B$
Uniform	$1-t^{-1}$ [a]	$1-\log^{-1}(t+1)$ [a]	$c-\log(t)+1$

Note. $\log(z)$ is the logarithm with base e. The constants are: γ is Euler's constant ($\simeq 0.577216$ with $\psi^{(0)}(1)=-\gamma$), $G=\frac{\Gamma(a+b)}{\Gamma(b)}$, where $\Gamma(z)$ is the gamma function, $A=-\psi^{(0)}(a)$, and $B=\psi^{(0)}(a+b)-\psi^{(0)}(a)$, where $\psi^{(0)}(z)$ is the digamma function.
[a] Limiting form for high t.
[b] Limiting form for high t that may have slow convergence for certain parameters.

Table A.1
Exact functions resulting from arithmetic averaging over the base functions with different mixing distributions on (learning or forgetting) rate parameter x.

Mixing distributions	Base functions ($\boldsymbol{t}$ is time, $\boldsymbol{x}$ is the learning rate)		
	$1-e^{-xt}$	$1-(1+t)^{-x}$	$c-\log(xt)$
Gamma	$1-(1+bt)^{-a}$	$1-(1+b\log(1+t))^{-a}$	$c-\log(bt)-\psi^{(0)}(a)$
Exponential	$1-(1+bt)^{-1}$	$1-(1+b\log(1+t))^{-1}$	$c-\log(bt)+\gamma$
Half-normal	$1-e^{\frac{\pi t^2}{4\theta^2}}\operatorname{erfc}\left(\frac{\sqrt{\pi}t}{2\theta}\right)$	$1-e^{\frac{\pi\log^2(1+t)}{4\theta^2}}\operatorname{erfc}\left(\frac{\sqrt{\pi}\log(1+t)}{2\theta}\right)$	$c-\log\left(\frac{\sqrt{\pi}}{2}\frac{1}{\theta}t\right)+\frac{\gamma}{2}$
Beta	$1-{}_1F_1(a;a+b;-t)$	$1-{}_1F_1(a;a+b;-\log(t+1))$	$c-\log(t)+B$
Uniform	$1-t^{-1}+\frac{e^{-t}}{t}$	$1-\frac{t}{(t+1)}\log(t+1)^{-1}$	$c-\log(t)+1$

Note. $\psi^{(0)}(z)$ is the digamma function, $\operatorname{erfc}(z)$ is the complementary error function, $\log(z)$ is the logarithm with base e, γ is Euler's constant ($\simeq 0.577216$ and $\psi^{(0)}(1)=-\gamma$), ${}_1F_1(a;b;z)$ is the Kummer confluent hypergeometric function, and $B=\psi^{(0)}(a+b)-\psi^{(0)}(a)$.

the expected time-course of learning or forgetting. I also concur with Sidman's warning (Iversen, 2021; Sidman, 1952) that attaching theoretical implications to the shape of an averaged curve may lead to unwarranted conclusions, where averaging artefacts over subjects and items form at the very least an important source of distortion, as is demonstrated in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

I would like to thank Richard Chechile for helpful suggestions and comments.

Appendix A. Mathematical analyses: Methods

Table A.1 shows the exact results of arithmetic averaging, which I will derive below for each of the three base functions in turn: exponential, power, and logarithm, for the three (learning or forgetting) rate distributions: gamma, half-normal, and beta. Two special cases, the exponential and uniform distribution, will be derived as well.

A.1. Arithmetic averaging

A.1.1. Exponential base functions

Both exact and limit results for the gamma, exponential, half-normal, and uniform distributions were already derived in Murre and Chessa (2011, p.596ff), and I will, therefore, only derive the result for beta distribution here. The averaged curve $p_A(t)$ can be derived by mixing the beta distribution on the rate parameter x of the base function through integration, as follows:

$$\begin{aligned}
p_A(t) &= \int_0^1 \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\left(1-e^{-xt}\right)dx \\
&= 1-\int_0^1 \frac{x^{a-1}(1-x)^{ffb-1}}{B(a,b)}e^{-xt}dx \\
&= 1-{}_1F_1(a;a+b;-t)
\end{aligned}$$

where ${}_1F_1(a;b+b;-t)$ is the Kummer confluent hypergeometric function and $B(a,b)$ is the beta function. This is a well-known result that can, for example, be found in Abramowitz and Stegun (1965, p. 505, Eq. 13.2.1), assuming that $a>0$ and $b>0$ are real valued parameters. The limiting form in Table 1 is based on Equations 13.5.1, 13.5.3, and 13.5.4 in the same chapter (p.508), which I will not repeat here in detail, only giving the final result:

$$\lim_{t\to\infty} {}_1F_1(a;a+b;-t) = \frac{\Gamma(a+b)}{\Gamma(b)}t^{-a}$$

More details on this derivation can also be found in Luo and Lin (2015, e.g., p. 5, Equation 1.30). Note that this convergence can be quite slow for certain values of a and b, whereas for others we get an exact solution that converges quite quickly to a power function. Notably for $a=1$ and $b=1$, we have a special case of the beta distribution, namely the standard uniform distribution, giving:

$${}_1F_1(1;2;-t) = \frac{1-e^{-t}}{t},$$

This is again a well-known identify, the result of which rapidly converges to t^{-1} for high t.

A more general form the uniform distribution has pdf

$$f(\mu)=\begin{cases}\frac{1}{b-a} & \text{for } a\le\mu\le b,\\ 0 & \text{for } \mu<a \text{ or } \mu>b.\end{cases}$$

If we integrate this function directly with the exponential individual learning curves, we obtain

$$P_A(t)=\int_a^b\frac{e^{-\mu t}}{b-a}d\mu=\frac{e^{-at}-e^{-bt}}{t(b-a)}.$$

If $b>a>0$ and if b approaches a very closely, we once more obtain an exponential function. This is to be expected because in that case (nearly) all subjects will have the same learning rate a:

$$\lim_{b\to a}\frac{e^{-at}-e^{-bt}}{t(b-a)}=e^{-at}.$$

We find a similar limit result with the averaged curve of the beta distribution (with bounds 0 and 1 as usual) when b is close to 0. This means that the mean $a/(a+b)$ would approach 1 and the variance $(ab)/((a+b)2(a+b+1))$ would approach 0 in the limit. The plot of such a beta distribution shows a thin pole close to 1. The limit is

$$\lim_{b\to 0}{}_1F_1(a;a+b;-t)=e^{-t}$$

This is based on Equation 13.6.12 in Abramowitz and Stegun (1965, p. 509). In other words, working with the beta distribution we also find that if most learners are close to the same high learning rate, the averaged function retains its original shape, as is to be expected. A variant where we have a strongly bimodal distribution is by taking both a and b small. If for simplicity we take $a=b$, we obtain

$$\lim_{a\to 0}{}_1F_1(a;a+a;-t)=\frac{1}{2}\left(1+e^{-t}\right)$$

This result simply reflects the fact that half the learners do not learn anything at all.

A.1.2. Power base functions

We first examine mixing with the gamma distribution:

$$1-\int_0^\infty\frac{\left(b^{-a}x^{a-1}e^{-\frac{x}{b}}\right)}{\Gamma(a)}(t+1)^{-x}\,dx=1-(1+b\log(t+1))^{-a}$$

Here, we can use the earlier result (Murre & Chessa, 2011) that

$$1-\int_0^\infty\frac{\left(b^{-a}x^{a-1}e^{-\frac{x}{b}}\right)}{\Gamma(a)}e^{-xt}dx=1-(1+bt)^{-a}$$

Considering that $(t+1)^{-x}=e^{\log\left((t+1)^{-x}\right)}=e^{-x\log(t+1)}$, we can change t to $\log(t)$ to obtain the desired result:

$$1-\int_0^\infty\frac{\left(b^{-a}x^{a-1}e^{-\frac{x}{b}}\right)}{\Gamma(a)}e^{-x\log(t+1)}dx=1-(1+b\log(t+1))^{-a}$$

This same approach can also be used for the half-normal and beta distributions, while noting that t itself does not take part in the integration, as we are integrating over x. Also, for the limiting forms we can do a simple substitution. As an example, I give the full derivation for the limiting form of the uniform distribution, which we can base on the special case shown above where we substitute $\log(t+1)$ for t.

$$\begin{aligned}{}_1F_1(1;2;-\log(t+1))&=\frac{t}{(t+1)\log(t+1)}\\&=\frac{1}{\log(t+1)}-\frac{1}{(t+1)\log(t+1)}\end{aligned}$$

This converges to the limiting form shown in Table 1:

$$\lim_{t\to\infty}\left(\frac{1}{\log(t+1)}-\frac{1}{(t+1)\log(t+1)}\right)=\frac{1}{\log(t+1)}$$

The power function as used here, $1-(t+1)^{-x}$, has as a limitation that it is not scale-free: changing the time units from say seconds to minutes would alter its shape. This can be remedied by using a slightly more complex form: $1-(yt+1)^{-x}$, where y allows adjusting to arbitrary time units. In the expressions above we can simply substitute yt for t to obtain this more general result. We could also consider a mixing distribution on y given that is also some type of rate parameter that could potentially differ between subjects. From a theoretical perspective, however, it makes more sense to consider this parameter to be fixed and shared between all subjects (see Murre (2014) for further discussion on subject-dependent versus shared parameters).

A.1.3. Logarithmic base functions

Estes (1956) demonstrates that a sufficient criterion for arithmetic averaging over base functions, while retaining their basic shape, is that the higher orders of the Taylor expansion, above 1, are zero, which is the case for the logarithm. As above, once we know the distribution of the learning (or forgetting) rate, we can also calculate the constants of the averaged curve. We will use the unconditioned logarithmic base function, $c-\log(tx)$, noting that it can only be considered as a viable function of learning or forgetting over short stretches of the time domain, unless the performance is measured in units that can become negative.

The integral to be evaluated has the general form:

$$\int_0^\infty f(x)(c-\log(tx))\,dx$$

where $f(x)$ is the pdf of a statistical distribution and the integration is over the domain of that distribution (here taken as 0 to ∞). This can be rearranged as:

$$c-\int_0^\infty(f(x)\log(t)+f(x)\log(x))\,dx$$

Here, we remark that

$$\int_0^\infty f(x)\log(t)\,dx=\log(t)\int_0^\infty f(x)\,dx=\log(t)$$

because by definition the integral over the domain of a pdf is always 1, and

$$\int_0^\infty f(x)\log(x)\,dx$$

is the logarithm of the geometric mean of $f(x)$.

So, we conclude that mixing a distribution with the rate of a logarithmic base function always gives

$$\int_0^\infty f(x)(c-\log(tx))\,dx=c-\log(t)-\log(\mathrm{GM}(f))$$

where $GM(f)$ is the geometric mean of the distribution.

As an example, we investigate mixing with the gamma distribution:

$$c-\int_0^\infty\frac{\left(b^{-a}x^{a-1}e^{-\frac{x}{b}}\right)}{\Gamma(a)}\log(tx)\,dx=c-\log(bt)-\psi^{(0)}(a)$$

We substitute bx for x, which gives

$$c-\int_0^\infty\frac{b^{-a}e^{-x}(bx)^{a-1}\log(btx)}{\Gamma(a)}b\,dx$$

where we use the first derivative of bx, which is $b\,dx$. This resolves to

$$c-\int_0^\infty\frac{e^{-x}x^{a-1}\log(btx)}{\Gamma(a)}dx$$

which can be rearranged as

$$c - \log(bt)\int_0^\infty \frac{e^{-x}x^{a-1}}{\Gamma(a)}dx - \int_0^\infty \frac{e^{-x}x^{a-1}}{\Gamma(a)}\log(x)\,dx$$
$$= c - \log(bt) - \psi^{(0)}(a)$$

where the first integral is over the gamma distribution with $b = 1$, giving 1, and the solution of the second integral is a standard result (e.g., Olver et al., 2020, Eq. 5.9.19, see https://dlmf.nist.gov/5.9#E19), adapted to fit the notation above. It is also the logarithm of the geometric mean of the gamma distribution.

For the half-normal distribution, the geometric mean cannot easily be found in existing sources, so I give the derivation of the averaged curve $P_A(t)$ in full.

$$\begin{aligned}
P_A(t) &= c - \int_0^\infty \frac{2\theta e^{-\frac{\theta^2x^2}{\pi}}\log(tx)}{\pi}dx \\
&= c - \log(t)\int_0^\infty \frac{2\theta e^{-\frac{\theta^2x^2}{\pi}}}{\pi}dx - \int_0^\infty \frac{2\theta e^{-\frac{\theta^2x^2}{\pi}}\log(x)}{\pi}dx \\
&= c - \log(t) - \int_0^\infty \frac{2\theta e^{-\frac{\theta^2x^2}{\pi}}\log(x)}{\pi}dx
\end{aligned}$$

Changing variables from x to $u = \frac{\theta x}{\sqrt{\pi}}$, with $dx = \frac{\sqrt{\pi}}{\theta}du$, gives

$$= c - \log(t) - \int_0^\infty \frac{\left(2\theta e^{-u^2}\right)\log\left(\frac{\sqrt{\pi}u}{\theta}\right)}{\pi}\frac{\sqrt{\pi}}{\theta}du$$
$$= c - \log(t) - \frac{2}{\sqrt{\pi}}\left(\int_0^\infty e^{-u^2}\log\left(\frac{\sqrt{\pi}}{\theta}\right)du\right.$$
$$\left.+\int_0^\infty e^{-u^2}\log(u)\,du\right)$$
$$= c - \log(t) - \frac{2}{\sqrt{\pi}}\left(\tfrac{1}{4}\sqrt{\pi}\left(2\log\left(\tfrac{1}{\theta}\right) + \log(\pi)\right)\right.$$
$$\left.-\tfrac{1}{4}\sqrt{\pi}\left(\gamma + \log(4)\right)\right)$$
$$= c - \log\left(\frac{\sqrt{\pi}t}{2\theta}\right) + \frac{\gamma}{2}$$

Mixing with the beta distribution can most easily be derived by taking the logarithm of its geometric mean – as outlined above – which is readily found (e.g., Vogel, 2020):

$$GM_{beta} = e^{\psi^{(0)}(a)-\psi^{(0)}(a+b)}$$

A great limitation of the simple logarithmic function used here is that it is not well-conditioned. I considered the function $\log(tx+1)/\log(Ux+1)$ as a more viable candidate (Chechile, 2022), but I was unable to derive any useful results from this. Another approach is to wrap the logarithm in a power function (with exponent -1): $(\log(tx+1)+1)^{-1}$. Though it is no longer a 'pure' logarithm, it is well-conditioned. Unfortunately, I could only find a useful result when mixing with the uniform distribution:

$$\frac{\mathrm{Ei}(\log(t+1)+1) - \mathrm{Ei}(1)}{et}$$

where Ei is the exponential integral. Taking into account that $E_i(-x) = -E_1(x)$, we use the well-known bracketing result:

$$\frac{1}{2}e^x\log\left(-\frac{2}{x}+1\right) < -\mathrm{Ei}(x) < e^x\log\left(-\frac{1}{x}+1\right)$$

On the basis of this, after substituting the correct terms in the left bound, we arrive at a somewhat more useful limit result for high t:

$$-\frac{1}{2}\log\left(1 - \frac{2}{\log(t+1)+1}\right)$$

This result is broadly in line with those obtained above, namely that averaging over a power base function tends to yield a logarithm function. I was unable to derive any useful results for the gamma, beta, or half-normal distribution.

A.2. Arithmetic averaging of shifted exponential and power base functions

We are interested in deriving the effect on arithmetic averaging with a mixing distribution $f(x)$ that has been shifted to the right with a distance of z. That is, the probability density below z is 0. In case of a learning curve this implies that there are no very slow learners. The derivation starts from changing the pdf of the distribution to $f(x-z)$. For the power base function, this would give the following expression for the averaged curve:

$$\int_z^\infty f(x-z)(t+1)^{-x}dx$$

To solve the integral, we use a change of variables, using $(t+1)^{-x} = (t+1)^{-(x-z)}(t+1)^{-z}$ to rewrite the product of the pdf with the power base function to

$$\int_z^\infty f(x-z)(t+1)^{-(x-z)}(t+1)^{-z}dx$$

Then, we change variable from $x - z$ to x:

$$(t+1)^{-z}\int_0^\infty f(x)(t+1)^{-x}dx = (t+1)^{-z}P_A(t)$$

Here $P_A(t)$ is the averaged curve using the not-shifted pdf. The same technique can be used with other distribution functions and the power base function. With the exponential, we can use the fact that $e^{-xt} = e^{-(x-z)t}e^{-zt}$. As above, we can take e^{-zt} out of the integration, and change variable to obtain

$$e^{-zt}P_A(t)$$

For other base functions, this usually does not work as elegantly, e.g., for $c - \log(tx - tz)$, it is not obvious how to remove the $\log(-tz)$ from the integral.

A.3. Geometric averaging of exponential and power base functions

It is well-known that geometric averaging over exponential and power base functions leaves the shape unaffected. For completeness, I still include the outline of the proof here.

For a mixing distribution $f(x)$ with (arithmetic) mean μ, assuming a domain of 0 to ∞, we calculate the averaged curve $P_A(t)$ as follows:

$$\begin{aligned}
P_A(t) &= \exp\left(\int_0^\infty \log\left(e^{-xt}\right)f(x)\,dx\right) \\
&= \exp\left(\int_0^\infty -txf(x)\,dx\right) \\
&= \exp\left(-t\int_0^\infty xf(x)\,dx\right) \\
&= e^{-\mu t}
\end{aligned}$$

For a power base function, the proof is analogous:

$$\begin{aligned}
P_A(t) &= \exp\left(\int_0^\infty \log\left((t+1)^{-x}\right)f(x)\,dx\right) \\
&= \exp\left(\log\left((t+1)^{-1}\right)\int_0^\infty xf(x)\,dx\right) \\
&= \exp\left(\log\left((t+1)^{-1}\right)\mu\right) \\
&= \exp(\log((t+1)^{-\mu})) \\
&= (t+1)^{-\mu}
\end{aligned}$$

For other types of base functions, e.g., logarithmic functions, such a simple proof cannot be obtained, and I found that most geometric means of mixing distributions could either not be calculated or yielded very complicated expressions that gave little additional insight. Also, the geometric averages of mirrored base functions or those with an asymptote did not yield a clear relation with the original base functions, at least not of the general nature above, where for exponential and power functions the shape is fully retained.

As an example, I give the geometric average of a mirrored exponential base function with a uniform rate distribution (assuming $t > 0$):

$$\begin{aligned} \mathrm{P}_A(\mathrm{t}) &= \exp\left(\int_0^1 log\,(1 - exp\,(-xt))\,1dx\right) \\ &= e^{\frac{-6\mathrm{Li}_2\left(e^t\right)-3t^2-6i\pi t+\pi^2}{6t}} \end{aligned}$$

where $\mathrm{Li}_2(z)$ is a polylogarithm. Though this function does converge to 1, it is not close to $1-\exp(-\mu t)$, where $\mu = \frac{1}{2}$, as we saw above (nor for any other values of μ). In other words, geometric averaging of a mirrored exponential function does not (always) retain the shape of the base function.

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical table*. Courier Dover Publications.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–730.

Bakan, D. (1954). A generalization of Sidman's results on group and individual functions, and a criterion. *Psychological Bulletin*, *51*(1), 63–64. http://dx.doi.org/10.1037/h0058163.

Brown, S. D., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers*, *35*(1), 11–21.

Chechile, R. A. (2018). *Analyzing memory: The formation, retention, and measurement of memory*. Cambridge, MA: MIT Press.

Chechile, R. A. (2022). Using hazard and surrogate functions for understanding memory and forgetting. *AppliedMath*, *2*(4), 518–546.

Ebbinghaus, H. (1880). *Urmanuskript "Ueber das gedächtniz"*. Passau: Passavia Universitätsverlag.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. New York: Dover.

Ebbinghaus, H. (1913/1885). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University, (H. A. Ruger & C. E. Bussenius, Trans.).

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. http://dx.doi.org/10.1037/h0045156.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Iversen, I. H. (2021). Sidman or statistics? *115*(1), 102–114. http://dx.doi.org/10.1002/jeab.660.

Luo, X. D., & Lin, W. C. (2015). Some new properties of confluent hypergeometric functions. arXiv preprint arXiv:.06465.

Murre, J. M. J. (2014). S-shaped learning curves. *Psychonomic Bulletin & Review*, *21*, 344–356. http://dx.doi.org/10.3758/s13423-013-0522-0.

Murre, J. M. J., & Chessa, A. G. (2011). Power laws from individual differences in learning and forgetting: Mathematical analyses. *Psychonomic Bulletin & Review*, *18*, 592–597. http://dx.doi.org/10.3758/s13423-011-0076-y.

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *Plos One*, *10*(7), Article e0120644. http://dx.doi.org/10.1371/journal.pone.0120644.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Olver, F. W. J., Olde Daalhuis, A. B., Lozier, D. W., Schneider, B. I., Boisvert, R. F., Clark, C. W., . . ., & McClain, M. A. (Eds.), (2020). *NIST digital library of mathematical functions*. (Release 1.1.4 of 2021-01-15. ed.). Retrieved from http://dlmf.nist.gov/.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, *49*(3), 263–269. http://dx.doi.org/10.1037/h0063643.

Vogel, R. M. (2020). The geometric mean? *Communications in Statistics. Theory and Methods*, *49*, 1–13. http://dx.doi.org/10.1080/03610926.2020.1743313.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory and Cognition*, *2*, 775–780.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.