Spurious Power Laws of Learning and Forgetting: Mathematical and Computational Analyses of Averaging Artifacts

Jaap M. J. Murre (Jaap@Murre.com)
Department of Psychology, Roetersstraat 15
1018 WB Amsterdam, The Netherlands

Antonio G. Chessa (AntonioGChessa@Yahoo.com)
Statistics Netherlands (CBS), P.O. Box 4000
2270 JM, Voorburg, The Netherlands

Abstract

It has frequently been claimed that learning performance improves with practice according to the so-called "Power Law of Learning". Similarly, forgetting may follow a Power Law. It has been shown on the basis of extensive simulations that such Power Laws may emerge as artifacts through averaging functions with other shapes. Here, we present a mathematical analysis that power functions will indeed emerge as a result of averaging over exponential functions, if the distribution of learning rates follows a gamma distribution. Power Laws may, thus, arise as a result of data aggregation over subjects or items. Through a number of simulations we further investigate to what extent these findings may affect empirical results in practice. We conclude that spurious Power Laws will be more likely with large numbers of subjects and shorter time scales and with gamma distributions with much probability mass close to zero and with a not too low variance.

Keywords: Learning; forgetting; Power Law; power function; exponential function; averaging; curve fitting

Power Laws of Learning and forgetting

It has frequently been claimed that learning performance P improves with practice time t according to the so-called Power Law of Learning or that forgetting of learned material follows a power function (e.g., J. R. Anderson & Schooler, 1991; Newell, 1981; Wixted, 2004). In its simplest form, a power function is simply a function of the shape $P = t^{\mu}$, where μ is the learning (or forgetting) rate and t is number of learning episodes or time. P may refer to how accurate or how fast we carry out a learned activity.

Power functions have the characteristic that the learning rate slows down with prolonged practice. The above equation is not correct if P denotes a probability, μ is negative and t is small. For example, for $t = 0.5$ and $\mu = -0.1$, we have $P = 0.5^{-0.1} = 1.072$. This would give a probability greater than 1, which is impossible. We can easily remedy this by adding 1 to t, thus obtaining $P = (t+1)^{\mu}$. This form ensures that its value remains properly scaled as a probability (i.e., remains between 0 and 1) if μ is negative.

Several authors dispute that learning follows a power function (e.g., Heathcote, Brown, & Mewhort, 2000), reporting exponential curves for individuals. Exponential curves have shape $P = \mu^{t}$. If learning shows an exponential improvement, the learning process itself does not slow down but continues at the same pace. These opposing viewpoints can be reconciled, if averaging over individual exponential curves would yield an averaged power function. This has indeed been found in an extensive simulation study (R. B. Anderson, 2001). The motivation by Anderson for carrying out a simulation study rather than a mathematical analysis was that a mathematical proof had not been established and may in fact be impossible. As we will demonstrate here, however, this is not the case. For at least one relevant case, a mathematical proof can be derived, which we will discuss below. Note that though we use the Power Law of Learning as a starting point of our analysis, our proof is general and applies to any situation where the assumptions are met. In particular, it also applies to the shape of forgetting functions.

Exponential learning curves

Our measure of learning performance is the probability $p(t)$ that a student will be correct on a certain test item (e.g., knowing foreign language vocabulary) after study time t. In the analysis, we will first assume an exponential learning curve for individual students: $P = p(t) = 1 - e^{-\mu t}$ with $\mu \geq 0$. Such a curve starts at zero performance at $t = 0$ and will reach an asymptote at 1 (100% correct) given enough study time. Our second assumption is that students' learning rates

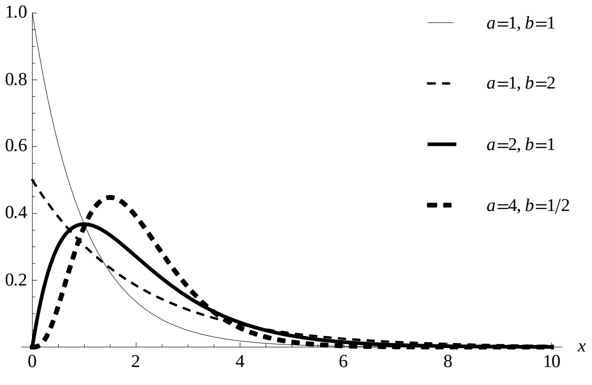

Figure 1: Illustration of the flexibility of the gamma distribution. Shown are plots for different values of *a* and *b*.

are not all equal to μ. Instead we make the more reasonable assumption that some will be fast learners (high μ) and others slow learners (low μ). Exactly how learning rates are distributed is an empirical question. Our aim here, however, is limited to showing that for at least one probability distribution the shape of the averaged curve can be derived. We will then investigate the implications of this in practice through simulations.

Gamma distributed learning rates

We will here consider the case where learning rates follow a *gamma distribution*. This is a well-known probability distribution that can take different shapes depending on its parameters *a* (the 'shape' parameter) and *b* (the 'scale' parameter). If the shape parameter *a* is 1, the gamma distribution becomes the *exponential distribution* as a special case. The mean of the gamma distribution is given by the product *ab* and the variance by ab^2. As can be seen from Figure 1, its shape is flexible and may vary from a peaked distribution where most learners tend to have low or average learning rates, to a broader distribution where learning rates are more variable.

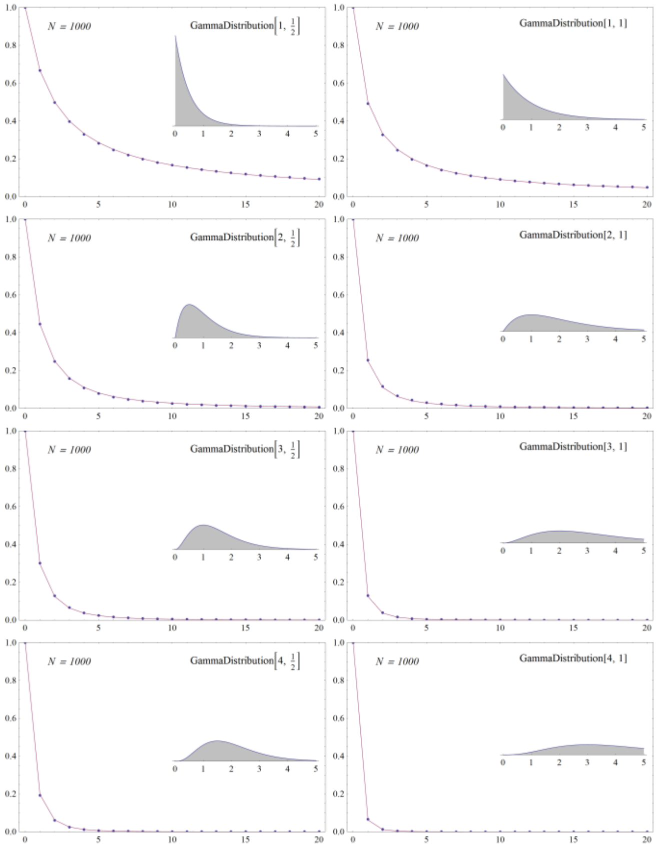

Figure 2: Convergence of the simulated exponential curves (dots) to predicted power curve (line) for different shapes of the learning rate distribution. The inserts give the shape of the distributions used to generate each plot.

Effects of averaging

In Appendix A, we show that, if we assume that the learning rates μ of individual subjects follow a gamma distribution, the average $p_A(t)$ of a number of exponential learning curves will approach $p_A(t) = 1 - (1 + a\,t)^{-b}$. This is a power function, which is properly scaled as a probability (i.e., remaining between 0 and 1) and starts at zero performance at $t = 0$.

In Figure 2, we present an illustration of this. For different shapes of the learning rate distribution we have generated 1000 artificial subjects with their learning rate drawn randomly from the distribution. Each subject learns according to an exponential curve. The different plots show the averaged curve over the 1000 exponential curves. We can observe that the simulated learning curves (dots) approach the theoretical power function (line) very closely.

Normally, experiments will have a number of subjects much smaller than 1000 and one might wonder whether the theoretical result still holds. In Figure 3, we show the distance (root of the squared differences, RSD) between the simulated and predicted curves for increasing numbers of subjects in an experiment that measured learning over 10 episodes. Each data point represents the average of 1000 simulated experiments. In larger experiments, with higher numbers of subjects, the simulated data continues to approach the theoretical data.

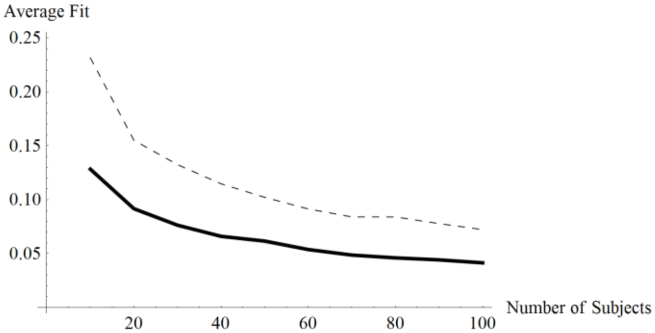

Figure 3: Experiments with larger number of subjects fit the power function increasingly well. The dashed line is based on a gamma distribution with (*a*, *b*) = (1, 0.5), the solid line (*a*, *b*) = (2, 0.5).

In the following series of simulations, we studied the effect of varying the parameters of the gamma distribution for a fixed study time (we used learning over 20 episodes). Each simulated experiment had 20 subjects, for which we calculated the average distance (RSD). Gamma distribution parameters *a* and *b* were varied from ¼ to 2 in steps of 0.05. For all data points shown in Figure 4, we took the average of 1000 simulated experiments.

As can be observed, the predicted power function is closer to the simulated data with lower *a* and higher *b* (up to a point). Lowering *a* shifts the mass of the distribution towards lower values (see plots in the left column of Figure 2, which shows shapes for increasing *a*). More slow learners

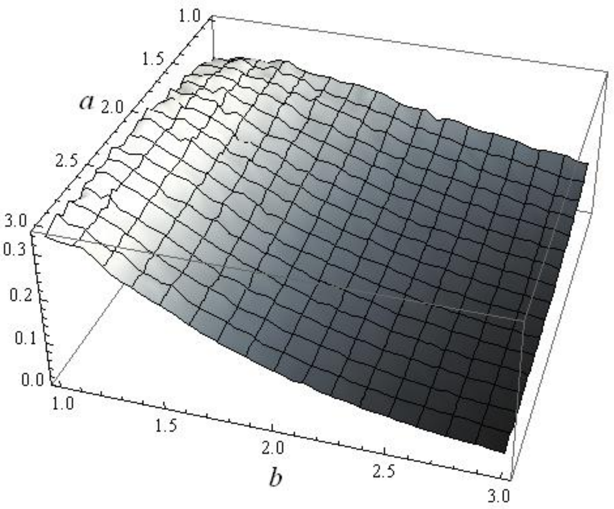

Figure 4: Distance between the theoretical and simulated curves as a function of gamma distribution parameters *a* and *b*. Each simulated experiment had 20 subjects and a time scale of 20 episodes (e.g., minutes or days).

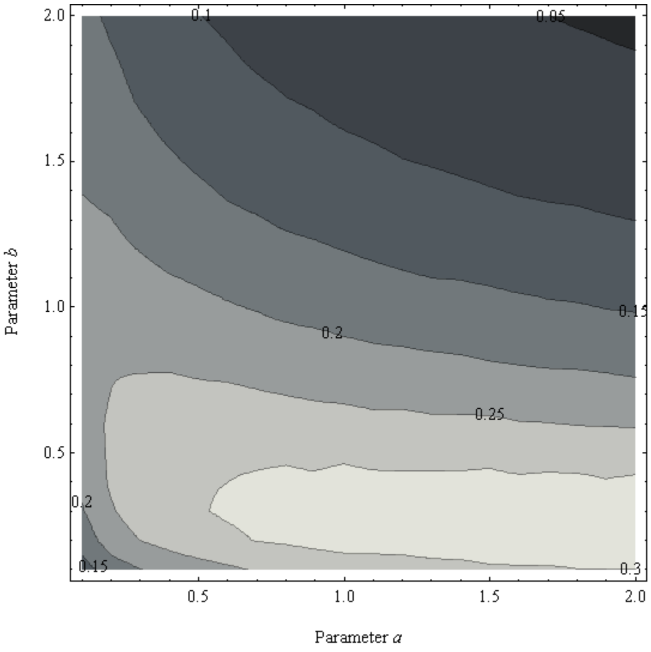

Figure 5: Contour plot of a similar simulation as Figure 4 but with *a* and *b* varied in steps of 0.1 and with each data point based on 10,000 simulated experiments.

will cause a 'thicker tail' in the learning curve, as it will approach 100% performance more slowly due to those learners that have low learning rates. Thick tails are seen as a characteristic of power functions. Increasing *b* increases the variance of the learning rate distribution. The opposite effect, decreasing the variance, will cause a narrower peak around the mean value. With very similar learning rates for all subjects, we would expect the averaged curve to also be quite similar to the individual learning curves, which we continue to assume to be exponential. Widening the peak will change the shape of the averaged curve from exponential to more power, as predicted by our derivation. In Figure 5 a contour plot is shown of a similar simulation, but now varying *a* and *b* from 1/10 to 2.0 in steps of 0.1. Each data point represents the average of 10000 simulated experiments with 20 subjects and 20 time episodes.

In the simulations thus far, we have compared the predicted power curve with the results of simulated experiments that vary in numbers of subjects, length of the learning curve, and shape of the learning rate distribution. One might well wonder, what would happen if we fitted both a power function and an exponential function to the averaged, exponential learning curves. To investigate this we simulated experiments with different rate distributions (drawn again from the gamma distribution) and with increasing numbers of subjects. Each simulated curve was fitted to an exponential function and a power function.

To enable simulations of thousands of experiments we used a linear model, fitted to the log-transformed data for the exponential function and the log-log transformed data for the power function. Though we are aware that this is not the most reliable method to find power laws, this method is often used in practice, e.g., by plotting the data with log transformed axes. With log-log axes, a power function will show up as a straight line. If only the ordinate is log-

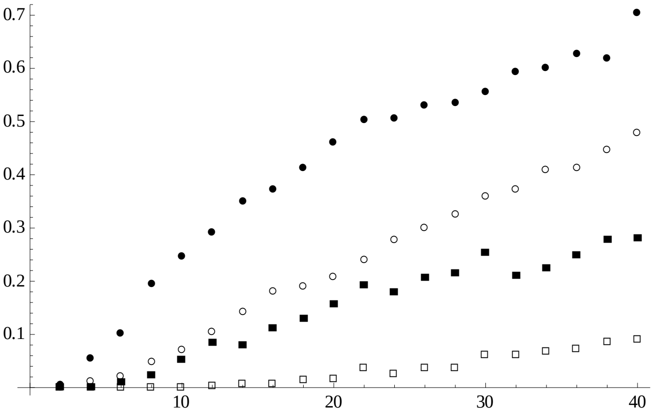

Figure 6: For simulated experiments with increasing numbers of subjects, the fraction is shown for which the power function gave a better fit on the Bayesian Information Criterion (a measure of goodness-of-fit), as opposed to an exponential function. Each point is the average of 1000 simulated experiments. Of the four curves, '●' has (*a*, *b*) = (2, 1) and episodes = 10, '■' has (*a*, *b*) = (4, 1/2) and episodes = 10, '○' has (*a*, *b*) = (2, 1) and episodes = 20, '□' has (*a*, *b*) = (4, 1/2) and episodes = 20.

transformed (i.e., data, but not time), an exponential function will give a straight line. The fit of the two function types was compared using the Bayesian Information Criterion (BIC), which is a popular goodness-of-fit criterion. For every simulation, the best fitting function was chosen. We were, thus, able to plot the fraction of times a power function was selected, which is shown in Figure 6 for (*a*, *b*) is (2, 1) and (4, ½) and for simulated experiments with time scales of 10 and 20 time units (e.g., seconds or days). The experiments in this simulation can be compared to those where an experimenter decides to fit empirical data to two different functions to see what fits best.

As can be seen in Figure 6, (*i*) averaging over more subjects increases the chances of finding a (spurious) fit to a power function, and (*ii*) extending the time scale in a curve decreases the chances of power function fits.

The BIC is a excellent measure of fit that takes into account relevant information such as number of free parameters. For many researchers, however, the proportion variance explained or r^2 is more meaningful. In Figure 7, we therefore show for one particular choice of learning rate distribution (a = 2, b = 1) and number of episodes (10) what the r^{-2} is. As can be observed, the variance explained is quite good for nearly all simulated experiments. For realistic numbers of subjects, the power function (open squares) gives r^2 values in excess of 0.985, even though these fits are an artifact caused by averaging exponential functions.

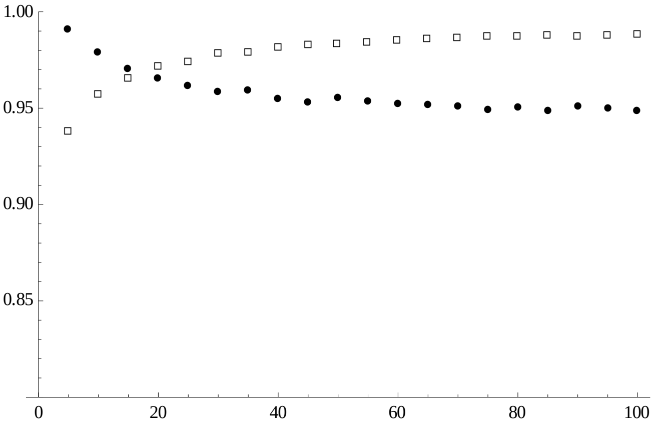

Figure 7: For simulated experiments with increasing numbers of artificial subjects, the r^2 value (fraction of variance explained) is shown for the fits of exponential function, '●', and power function '□'. Each point is the average of 1000 simulated experiments with gamma distribution parameters (a, b) = (2, 1) and number of learning episodes = 10.

Discussion

We presented a mathematical analysis of the effects of averaging exponential curves, demonstrating that averaging exponentials gives rise to spurious Power Laws, if the subjects have different learning rates that follow a gamma distribution. Through computer simulations, we also showed that (*i*) averaging over higher numbers subjects increases the chances of finding a good fit to a power function, and (*ii*) extending the time scale in the experiment decreases the chances of power function fits. We also showed that with gamma distribution parameter values that put the mass of distribution closer to zero or that turn it from narrow peak into a somewhat flatter distribution, the simulated data converge quickly to the predicted power curve.

Our analyses are based on assumptions that may be satisfied frequently, given the flexibility of the gamma distribution, offering a plausible explanation for the ubiquitous reports of the Power Laws of Learning and Forgetting. While this does not rule out that other processes may give rise to power laws (Wixted, 2004), we have adduced mathematical proof that power laws may arise as a result of mere data aggregation without reflecting directly the properties of fundamental cognitive processes, which may well be exponential in nature.

The theoretical result can be generalized to forgetting functions where we consider t to be time since the completion of learning. Using $p(t) = e^{-\mu t}$ for individual curves and assuming a gamma distribution of the individual forgetting rates μ, we obtain for the averaged forgetting curve: $p_A(t) = (1 + a\,t)^{-b}$. Our result is corroborated by a recent analysis (Lee, 2004) of over 200 forgetting studies taken from the literature, most of which average across participants. These forgetting curves are best modeled by the hyperbolic function 1/(1+*t*), which is a power function with exponent b = 1. If our analysis would apply to this case and the hyperbolic function does indeed emerge from averaging over exponential forgetting functions, we expect the distribution of the forgetting rates to be $f(\mu) \propto a^{-1} e^{-a^{-1}\mu}$. This i**s** an exponential distribution, implying that in these experiments we should observe rather many students that show little or no forgetting. This is not implausible with the short retention intervals often encountered in the psychology laboratory where there may not be enough time to allow sufficient forgetting for many subjects. The resulting ceiling effects would foster spurious Power Laws in the averaged forgetting curves.

The analysis can also be applied to averaging over items rather than students. We then assume that single items to be learned (e.g., foreign language words) have different learning rates, according to a gamma distribution. The learning curve averaged over items will then appear as a power function. Thus, even a single student may show a power learning curve based on averaged performance of heterogeneous items to-be-learned.

The analysis can be carried even further, to the level below that of a single item, namely to the features that make up its representation. Suppose, an item's representation can be approximated by discrete, uncorrelated elements and that the individual elements are learned according to an exponential learning function where the learning rates follow a gamma distribution. Then, if the item's strength during the learning process reflects the average strength of its features, the learning of such a single item will follow the power law of learning. A single item in this example may be of any type or modality (e.g., a cognitive representation of face or word meaning, but also of a complex activity such as a certain movement pattern in sports or music); as long as the assumption regarding its features are met, it will exhibit power law learning. The same applies to forgetting of single items. The effects of averaging at different levels will transform exponentially shaped curves into a power curve.

Acknowledgments

This research was funded by the Netherlands Organisation for Scientific Research.

References

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

Anderson, R. B. (2001). The power law as an emergent property. *Memory & Cognition, 7*, 1061-1068.

Chessa, A. G., & Murre, J. M. J. (2006). Modelling memory processes and Internet response times: Weibull or power-law? *Physica A, 366*, 539-551.

Feller, W. (1966). *An introduction to probability theory and its applications, Volume 2*. New York: Wiley.

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.

Lee, M. L. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology, 48*, 310-321.

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.

Wixted, J. T. (2004). On common ground: Jost's (1897) Law of Forgetting and Ribot's (1881) Law of Retrograde Amnesia. *Psychological Review, 111*, 864-879.

Appendix A

Here, we present the mathematical details of averaged learning curves. Individuals differ in learning rates. Exactly how individuals differ is described by a probability density function. We will couch the example in terms of learning processes but it should be pointed out that the results presented here are general and apply to forgetting and other processes as well.

The starting point of our analysis is the form of the function for the probability of correctly recalling a learned item for an individual subject. Recall takes place after cueing and retrieving information about an item. The success of recall will improve if learning time is increased or learning is enhanced.

Let us denote the initial amount of learned information stored per unit of practice time t by μ. Next, we denote the extent to which information will be recalled at time z since learning by a function $r(z)$, such that $r(0) = 1$, where $z = 0$ denotes the end of a learning trial. In a recent study, we derived and tested a recall function within an extreme-value theoretical framework (Chessa & Murre, 2006). Based on this study, we take the following general form for the recall function $p(t, z)$ of an individual subject at retention lag z, after learning time t:

$$p(t,z) = 1 - e^{-\mu\, t\, r(z)}. \qquad (1)$$

Note, that when $z = 0$, we have the special case of a simple exponential function of the shape $p(t) = 1 - e^{-\mu\, t}$ as discussed above. To simplify the analysis somewhat we will present most analysis in terms of averaging over $p(t) = e^{-\mu\, t}$.

Gamma distribution

We assume that individual learning rates μ follow a gamma distribution with density function

$$f(\mu) = \tfrac{1}{\Gamma(b)a^b}\, \mu^{b-1} e^{-\mu/a}, \qquad (2)$$

with parameters a, $b > 0$, where $\Gamma(b)$ is the well-known gamma function that equals $(b - 1)!$ for integer values of b.

Exponential functions

The recall function aggregated over subjects that learn exponentially but with different learning rates, which we denote as $p_A(t, z)$, is equal to the mathematical expectation of function (1) with respect to μ, that is:

$$\begin{aligned} p_A(t,z) &= \int_0^\infty p(t,z)\, f(\mu)\, \mathrm{d}\mu \\ &= \int_0^\infty \left(1 - e^{-\mu\, t\, r(z)}\right) \tfrac{1}{\Gamma(b)a^b}\, \mu^{b-1} e^{-\mu/a} \mathrm{d}\mu \qquad (3) \\ &= 1 - \left(1 + a\, t\, r(z)\right)^{-b}. \end{aligned}$$

where a and b originate from the gamma distribution. This function can be calculated by applying elementary probability calculus and is a standard result in statistics (e.g., Feller, 1966, p. 48), though it appears to be largely unknown in psychology. For $z = 0$, aggregate recall is equal to $1 - (1 + a\, t)^{-b}$ for practice only, which is the well-known 'Power Law of Practice'. The proof also applies to forgetting rates, e.g., when averaging over exponential forgetting curves with shape $p(t) = e^{-\mu\, t}$.