Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Classical statistical mechanics, equilibrium and non-equilibrium

An elementary renormalization-group approach to the generalized central limit theorem and extreme value distributions

Ariel Amir

John A. Paulsson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, United States of America
E-mail: arielamir@seas.harvard.edu

Received 27 August 2019
Accepted for publication 4 November 2019
Published 28 January 2020

Online at stacks.iop.org/JSTAT/2020/013214
https://doi.org/10.1088/1742-5468/ab5b8c

Abstract. The generalized central limit theorem is a remarkable generalization of the central limit theorem, showing that the sum of a large number of independent, identically-distributed (i.i.d) random variables with infinite variance may converge under appropriate scaling to a distribution belonging to a special family known as Lévy stable distributions. Similarly, the *maximum* of i.i.d. variables may converge to a distribution belonging to one of three universality classes (Gumbel, Weibull and Fréchet). Here, we rederive these known results following a mathematically non-rigorous yet highly transparent renormalization-group-inspired approach that captures both of these universal results following a nearly identical procedure.

Keywords: exact results, extreme value statistics, renormalisation group

J. Stat. Mech. (2020) 013214

 1742-5468/20/013214+18$33.00

Contents

1. Introduction

Consider some distribution $P(x)$ from which we draw independent random variables $x_1, x_2, \ldots, x_n$. If the distribution has a finite standard deviation σ and mean $\langle x\rangle$, we can define:

$$\xi \equiv \frac{\sum_{i=1}^{N}(x_i - \langle x\rangle)}{\sigma\sqrt{N}}, \tag{1}$$

and the central limit theorem (CLT) tells us that the distribution of ξ, $p(\xi)$, approaches a Gaussian with vanishing mean and a standard deviation of 1 as $N \to \infty$. What happens when $P(x)$ does *not* have a finite variance? Or a finite mean? Perhaps surprisingly, in this case the generalized central limit theoreom (GCLT) tells us that the limiting distribution belongs to a particular family (Lévy stable distributions), of which the Gaussian distribution is a proud member albeit *e pluribus unum*. Moreover, the familiar $\sqrt{N}$ scaling of the above equation does not hold in general, and its substitute will generally sensitively depend on the form of the *tail* of the distribution.

The results are particularly intriguing in the case of heavy-tailed distributions where the *mean* diverges. In that case the sum of N variables will be dominated by rare events, regardless of how large N is! Figure 1(c) shows one such example, where a running sum of variables drawn from a distribution whose tail falls off as $p(x) \sim 1/x^{3/2}$ was used. The code which generates this figure is remarkably simple, and included in the appendix. The underlying reason for this peculiar result is that for distributions with a power-law tail $p(x) \propto 1/x^{1+\mu}$, with $\mu \leqslant 1$, the distributions of both the *sum* and *maximum* of the N variables scale in the same way with N, namely as $N^{1/\mu}$—dramatically different from the $\sqrt{N}$ scaling we are used to from the CLT. (Note that we are excluding the regime $1 < \mu < 2$ here, since in that case if the mean $\langle x \rangle$ of the distribution is non-zero, the dominant term in the running sum will be $N\langle x \rangle$ rather than $N^{1/\mu}$.) The distribution of the *maximum* is known as the *extreme value distribution* or EVD (since for large N it inherently deals with rare, atypical events among the N i.i.d variables). Surprisingly, also for this quantity universal statements can be made, and when appropriately scaled this random variable also converges to one of three universality classes—depending on the nature of the tails of the original distribution from which the i.i.d variables are drawn.

Here, we will provide a straightforward derivation of these results. Although compact and elementary, to the best of our knowledge it has not been utilized previously, and is distinct (and simpler) than other renormalization-group approaches to the GCLT and to EVD, which are discussed later on. The derivation will not be mathematically rigorous—in fact, we will not even specify the precise conditions for the theorems to hold, or make precise statements about convergence. In this sense the derivation may be considered as 'exact but not rigorous', targeting a physics rather than mathematics audience ([1], for example, provides a rigorous treatment of many of the results derived in this paper). Throughout, we will assume sufficiently smooth probability distributions (what mathematicians refer to as probability density functions), potentially with a power-law tail such that the variance or mean may diverge (known as a 'fat' or 'heavy' tail).

1.1. Example: Cauchy distribution

Consider the following distribution, known as the Cauchy distribution:

$$p(x) = \frac{1}{\gamma\pi(1 + (\frac{x}{\gamma})^2)}. \tag{2}$$

Its characteristic function $\varphi(\omega) \equiv \int_{-\infty}^{\infty} p(x)\mathrm{e}^{\mathrm{i}\omega x}\mathrm{d}x$ is:

$$\varphi(\omega) = \mathrm{e}^{-\gamma|\omega|}. \tag{3}$$

Thus the characteristic function of a sum of N such variables is:

$$\varphi_N(\omega) = \mathrm{e}^{-N\gamma|\omega|}, \tag{4}$$

and taking the inverse Fourier transform we find that the distribution of the sum, $p_N(x)$, is *also* a Cauchy distribution:

$$p_N(x) = \frac{1}{N\gamma\pi(1 + (\frac{x}{N\gamma})^2)}. \tag{5}$$

Thus, the sum does not converge to a Gaussian, but rather retains its Lorentzian form. Moreover, it is interesting to note that the scaling form governing the width of the Lorentzian evolves with N in a different way than the Gaussian scenario: while in the latter the variance increases linearly with N hence the width increases as $\sqrt{N}$, here

J. Stat. Mech. (2020) 013214

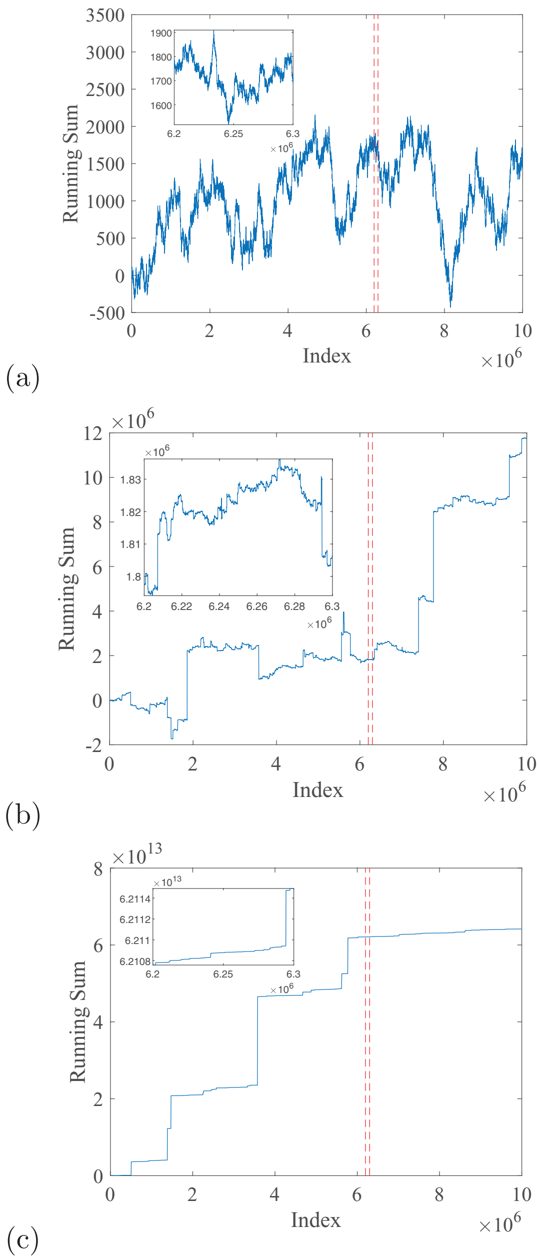

Figure 1. Running sum of independent, identically-distributed variables drawn from three distributions: Gaussian (top), Cauchy (middle) and a distribution with positive support and a power-law tail $1/x^{3/2}$ (bottom). See Appendix for details of the code. The insets illustrate the self-similar nature of the running sum, zooming into the small region of the original plot between the two vertical, dashed lines.

J. Stat. Mech. (2020) 013214

the scaling factor is linear in N. This remarkable property is in fact useful for certain computer science algorithms [2].

2. Self-similarity of running sums

To generate figure 1, we generate a set of i.i.d variables from a given distribution (Gaussian, Cauchy and a heavy-tailed distribution with positive support whose tail falls off as $1/x^{3/2}$). For each long sequence of random variables, the running sum is plotted. For the Gaussian case (or any case where the variance is finite), the result is the familiar process of diffusion: the variance increases linearly with 'time' (i.e. the index of the running sum). For figure 1(a) the mean vanishes, hence the running sum follows this random walk behavior. If we were to repeat this simulation many times, the result of the running sum at time N, a random variable of course, is such that when scaled by $1/\sqrt{N}$ it would follow a normal distribution with variance 1, as noted in equation (1). Another important property is that 'zooming' into the running sum (see the figure inset) looks identical to the original figure—as long as we do not zoom in too far as to reveal the granularity of the data. Figure 1(b) shows the same analysis for the Cauchy distribution of equation (2). As we have seen, now the scaling is linear in N. Nevertheless, zooming into the data still retains its Cauchy statistics. The mathematical procedure we will shortly follow to find *all* Lévy stable distributions will rely on this self-similarity. Indeed, assume that a sum of variables from some distribution converges—upon appropriate linear scaling—to some Lévy stable distribution. Zooming further 'out' corresponds to generating sums of Lévy stable variables, hence it retains its statistics. A dramatic manifestation of this is shown in figure 1(c). The initial distribution is *not* a Lévy stable, but happens to have a very fat tail, possessing infinite mean and variance. The statistics of the running sum converges to a Lévy stable distribution—in this case fortuitously expressible in closed form, corresponding to the Lévy distribution we will discuss in equation (28). Importantly, zooming into the running sum still retains its statistics, which in this case happens to manifest large jumps associated with the phenomenon of Lévy flights, which will be elucidated by our later analysis. Due to the self-similar nature, zooming into what seems to be flat regions in the graph shows that their statistical structure is the same, and they also exhibit these massive jumps. We also note that this renormalization-group-inspired idea has been utilized in the context of the 'conventional' CLT in [3] (Exercise 12.11). Notably, [4] discusses the deep connections between the CLT and RG approaches, and emphasizes the notion and relevance of 'Self-similar random fields', highly related to the self-similarity discussed here that forms the basis of our RG-inspired analysis. It should be emphasized that our approach is not an RG one *par excellence* and appears to be simpler than other RG approaches previously utilized in the context of the GCLT [5], as we elaborate on later.

3. RG-inspired approach to the generalized central limit theorem

We will look for distributions which are *stable*: this means that if we add two (or more) variables drawn from this distribution, the distribution of the sum will retain the same shape—i.e. it will be identical up to a potential shift and scaling, by some yet undetermined factors. If the sum of a large number of variables drawn from *any* distribution converges to a distribution with a well defined shape, it must be such a stable distribution. The family of such distributions is known as *Lévy stable*.

We shall now use an approach similar to those used in the context of RG (renormalization group) to find the general form of such distributions, which will turn out to have a simple representation in Fourier rather than real space—essentially because the characteristic function is the natural object to deal with here.

The essence of the approach relies on the fact that if we sum a large number of variables, and the sum converges to a stable distribution, then by definition taking a sum involving, say, twice the number of variables, will *also* converge to the same distribution—up to a potential shift and rescaling. This is illustrated visually in figure 1 by plotting the running sum of independently and identically distributed variables.

Defining the partial sums by s_n, the general (linear) scaling one may consider is:

$$\xi_n = \frac{s_n - b_n}{a_n}. \tag{6}$$

Here, a_n determines the width of the distribution, and b_n is a shift. If the distribution has a finite mean it seems plausible that we should center it by choosing $b_n = \langle x \rangle n$. We will show that this is indeed the case, and that if its mean is infinite we can set $b_n = 0$.

The scaling we are seeking is of the form of equation (6), and our hope is that if the distribution of ξ_n is $p_{\xi_n}(x)$, then:

$$\lim_{n\to\infty} p_{\xi_n}(x) = p(x) \tag{7}$$

exists, i.e. the scaled sum converges to some distribution $p(x)$, which is not necessarily Gaussian (or symmetric).

Let us denote the characteristic function of the scaled variable by $\varphi(\omega)$ (assumed to be approximately independent of n for large n). Consider the variable $y_n = \xi_n a_n$. Its distribution, p_{y_n} is:

$$p_{y_n}(y_n) \approx \frac{1}{a_n} p(y_n/a_n), \tag{8}$$

with p the limiting distribution of equation (7) (and the factor $\frac{1}{a_n}$ arising from the Jacobian of the transformation). The characteristic function of the variable y_n is:

$$\varphi_{y_n}(\omega) \approx \varphi(a_n \omega). \tag{9}$$

Consider next the distribution of the sum s_n. We have $s_n = y_n + b_n$, and its distribution, p_{s_n}, is:

$$p_{s_n}(s_n) = p_{y_n}(s_n - b_n). \tag{10}$$

Shifting a distribution by b_n implies multiplying the characteristic function by $e^{i\omega b_n}$. Therefore the characteristic function of the sum is:

$$\varphi_{s_n}(\omega) = \mathrm{e}^{\mathrm{i}b_n\omega}\varphi_{y_n}(\omega) \approx \mathrm{e}^{\mathrm{i}b_n\omega}\varphi(a_n\omega). \tag{11}$$

This form will be the basis for the rest of the derivation, where we emphasize our assumption that the characteristic function φ is n-independent.

Consider $N = n \cdot m$, where n, m are two large numbers. The important insight is to realize that one may compute s_N in two ways: as the sum of N of the original variables, or as the sum of m variables, each one being the sum of n of the original variables. The characteristic function of the sum of n variables drawn from the original distribution is given by equation (11). If we take a sum of m variables drawn from *that* distribution (i.e. the one corresponding to the sums of n's), then its characteristic function will be on the one hand:

$$\varphi_{s_N}(\omega) \approx \mathrm{e}^{imb_n\omega}(\varphi(a_n\omega))^m, \tag{12}$$

and on the other hand it is the distribution of $n \cdot m = N$ variables drawn from the original distribution, and hence does not depend on n or m separately but only on their product N. Therefore, assuming that n is sufficiently large such that we may treat it as a continuous variable, we have:

$$\frac{\partial}{\partial n}\mathrm{e}^{\mathrm{i}\frac{N}{n}b_n\omega+\frac{N}{n}log[\varphi(a_n\omega)]} = 0. \tag{13}$$

Defining $d_n \equiv \frac{b_n}{n}$, we find:

$$\Rightarrow iN\omega\frac{\partial d_n}{\partial n} - \frac{N}{n^2}\log(\varphi) + \frac{N}{n}\frac{\varphi'}{\varphi}\frac{\partial a_n}{\partial n}\omega = 0. \tag{14}$$

$$\Rightarrow \frac{\varphi'(a_n\omega)\omega}{\varphi(a_n\omega)} = \frac{\log(\varphi(a_n\omega))}{n\frac{\partial a_n}{\partial n}} - i\omega\frac{\partial d_n}{\partial n}\frac{n}{\frac{\partial a_n}{\partial n}}. \tag{15}$$

Multiplying both sides by a_n and defining $\tilde{\omega} \equiv a_n\omega$, we find that:

$$\frac{\varphi'(\tilde{\omega})\tilde{\omega}}{\varphi(\tilde{\omega})} - \log(\varphi(\tilde{\omega}))\frac{a_n}{n\frac{\partial a_n}{\partial n}} + i\tilde{\omega}\frac{\partial d_n}{\partial n}\frac{n}{\frac{\partial a_n}{\partial n}} = 0. \tag{16}$$

Since this equation should hold (with the same function $\varphi(\tilde{\omega})$) as we vary n, we expect that $\frac{a_n}{n\frac{\partial a_n}{\partial n}}$ and $\frac{\partial d_n}{\partial n}\frac{n}{\frac{\partial a_n}{\partial n}}$ should be nearly independent of n for large values of n. The equation for $\varphi(\tilde{\omega})$ then takes the following mathematical structure:

$$\frac{\varphi'}{\varphi} - \frac{C_1\log(\varphi(\tilde{\omega}))}{\tilde{\omega}} = iC_2, \tag{17}$$

with C_1, C_2 (real) constants. We may rewrite it using $u(\tilde{\omega}) \equiv \log(\varphi(\tilde{\omega}))$ as:

$$u' - \frac{C_1}{\tilde{\omega}}u = iC_2. \tag{18}$$

It is straightforward to solve the ODE and find that for $C_1 \neq 1$ its general solution is:

$$u(z) = A|\tilde{\omega}|^{C_1} + \frac{iC_2}{1-C_1}\tilde{\omega}, \tag{19}$$

while for $C_1 = 1$ the general solution is:

$$u(z) = A|\tilde{\omega}| + iC_2\tilde{\omega}\log(|\tilde{\omega}|), \tag{20}$$

with A in both equations an arbitrary complex constant. Importantly, note that since equation (17) is ill-defined at $\omega = 0$, the constant A may change as ω changes sign—we will shortly see that this indeed must be the case.

Going back to equation (16), we can also get the approximate scaling for the coefficients:

$$\frac{a_n}{n\frac{\partial a_n}{\partial n}} \approx C_1 \Rightarrow C_1\frac{\partial \log(a_n)}{\partial n} \approx 1/n. \tag{21}$$

This implies that:

$$\log(a_n) \approx \frac{1}{C_1}\log(n) + \text{constant} \Rightarrow a_n \propto n^{1/C_1}. \tag{22}$$

Similarly:

$$\frac{\partial d_n}{\partial n}\frac{n}{\frac{\partial a_n}{\partial n}} = C_2. \tag{23}$$

Hence:

$$\frac{\partial d_n}{\partial n} \propto n^{1/C_1-2}. \tag{24}$$

Therefore:

$$d_n = C_3 n^{1/C_1-1} + C_4 \Rightarrow b_n = C_3 n^{1/C_1} + C_4 n, \tag{25}$$

where we relied on our previous definition $d_n = b_n/n$. The first term in the formula for b_n will become a constant when we divide by the term $a_n \propto n^{1/C_1}$ of equation (6), leading to a simple shift of the resulting distribution. Upon dividing by the term a_n, the second term will vanish for large n when $C_1 < 1$. The case $C_1 > 1$ corresponds to the case of a variable with finite mean, in which case the $C_4 n$ term will be associated with centering of the scaled variable by subtracting their mean, as in the standard CLT.

A word of caution. The constraint imposed by our RG-inspired approach is insufficient in pinning down the scaling factor a_n precisely. Really, all we know is that $\lim_{n\to\infty}\frac{a_n}{n\frac{\partial a_n}{\partial n}}$ should tend to a constant. In the above, we solved the ODE resulting from equating this term to a constant, but it is easy to see that modulating this power-law by, e.g. logarithmic corrections (or powers thereof) would also satisfy the RG requirement. Similarly care should be taken in interpreting the power-law scaling of the coefficients b_n, as well as their counterparts in the 'extreme value distributions' later on. Nevertheless, in many applications knowing the leading order dependence of the coefficients on n suffices, which is adequately captured by the RG-inspired approach.

4. General formula for the characteristic function

According to equations (19) and (20), the general formula for the characteristic function of $p(\xi_n)$ for $C_1 \neq 1$ is:

$$\varphi(\omega) = \mathrm{e}^{A|\omega|^{C_1}+iD\omega} \tag{26}$$

the iD term is associated with a trivial shift of the distribution (related to the linear scaling of b_n) and can be eliminated. We will therefore not consider it in the following. The case of $C_1 = 1$ will be considered in the next section.

The requirement that the inverse Fourier transform of φ is a probability distribution imposes that $\varphi(-\omega) = \varphi^*(\omega)$. Therefore the characteristic function takes the form:

$$\varphi = \begin{cases} \mathrm{e}^{A\omega^{C_1}} & \omega > 0 \\ \mathrm{e}^{A^*|\omega|^{C_1}} & \omega < 0. \end{cases} \tag{27}$$

(As noted previously, the value of A in equation (26) was indeed 'allowed' to change at $\omega = 0$).

This may be rewritten as:

$$\varphi = \mathrm{e}^{-a|\omega|^{\mu}[1-i\beta sign(\omega) tan(\frac{\pi\mu}{2})]}, \tag{28}$$

where clearly $\mu = C_1$. The asymmetry of the distribution is determined by β. For this representation of φ, we will now show that $-1 \leqslant \beta \leqslant 1$, that $\beta = 1$ ($\beta = -1$) corresponds to a distribution with positive (negative) support, and $\beta = 0$ corresponds to a symmetric distribution.

Consider $p(x)$ which decays, for $x > x^*$, as

$$p(x) = \frac{A_+}{x^{1+\mu}}, \tag{29}$$

with $0 < \mu < 1$, and similarly has a left tail (at sufficiently negative x):

$$p(x) = \frac{A_-}{x^{1+\mu}}. \tag{30}$$

Generally, the form of the tails of the distribution dictate the form of the characteristic function for small ω—these are broadly referred to as Tauberian theorems [1]. For small, positive ω, a Tauberian theorem tells us that the Fourier transform is well approximated by:

$$\varphi(\omega) \approx 1 - \tilde{C}\omega^{\mu}(A_+\mathrm{e}^{-\mathrm{i}\mu\frac{\pi}{2}} + A_-\mathrm{e}^{\mathrm{i}\mu\frac{\pi}{2}}), \tag{31}$$

with $\tilde{C} \equiv \frac{\Gamma(1-\mu)}{\mu}$ (see [6] for a simple, non-rigorous derivation).

We can write this as $\varphi(\omega) \approx 1 - C\omega^{\mu}$, where now we have:

$$\frac{\mathrm{Im}(C)}{\mathrm{Re}(C)} = \frac{-sin(\frac{\pi}{2}\mu)}{cos(\frac{\pi}{2}\mu)}\left(\frac{A_+ - A_-}{A_+ + A_-}\right) = -tan(\frac{\pi}{2}\mu)\beta, \tag{32}$$

with β defined as:

$$\beta = \frac{A_+ - A_-}{A_+ + A_-}. \tag{33}$$

This clarifies the notation of equation (28), and why β is restricted to the range $[-1, 1]$.

A tail of tales—and black swans. It is interesting to note that unlike the case of finite variance, here the limiting distribution depends only on A_+ and A_-: the tails of the original distribution. The behavior is only dominated by these tails—even if the power-law behavior only sets in at large values of $|x|$! This also brings us to concept of a 'black swan': scenarios in which rare events—the probability of which is determined by the tails of the distribution—may have dramatic consequences. Here, such events dominate the sums. For a popular discussion of black swans and their significance, see [7].

In the case $1 < \mu < 2$, upon performing the shift of equation (6) the linear in ω contribution to the characteristic function near the origin will be eliminated, and a Tauberian theorem ensures that to leading order equation (31) still holds.

Special Cases

$\mu = 1/2$, $\beta = 1$: Lévy distribution

Consider the Lévy distribution:

$$p(x) = \sqrt{\frac{C}{2\pi}} \frac{\mathrm{e}^{-\frac{C}{2x}}}{(x)^{3/2}} \quad (x \geqslant 0). \tag{34}$$

The Fourier transform of $p(x)$ for $\omega > 0$ is

$$\varphi(\omega) = \mathrm{e}^{-\sqrt{-2iC\omega}}, \tag{35}$$

which indeed correspond to $\tan(\frac{\pi}{2}\frac{1}{2}) = 1 \rightarrow \beta = 1$.

$\mu = 1$: Cauchy distribution and more

The case $\mu = 1$, $\beta = 0$ corresponds to the Cauchy distribution. In the general case $\mu = 1$ and $\beta \neq 0$, we have seen that the general form of the characteristic function is, according to equation (20):

$$\varphi(\omega) = \mathrm{e}^{A|\omega| + iD\omega \log|\omega|}. \tag{36}$$

Relying on a similar Tauberian theorem to the previously quoted ones leads to:

$$\varphi(\omega) = \mathrm{e}^{-|C\omega|[1 - i\beta Sign(\omega)\phi]}; \quad \phi = -\frac{2}{\pi} log|\omega|, \tag{37}$$

where once again β is confined to the range $[-1, 1]$. This is the only exception to the form of equation (28).

5. RG-inspired approach for extreme value distributions

Consider the maximum of n variables drawn from some distribution $p(x)$, characterized by a cumulative distribution $C(x)$ (i.e. $C(x)$ is the probability for the variable to be smaller than x). It vanishes for $x \to -\infty$ and approaches 1 as $x \to \infty$. We will now find the behavior of the maximum for large n, that will turn out to also follow universal statistics—much like in the case of the GCLT—that depend on the tails of $p(x)$. This was discovered by Fisher and Tippett, motivated by an attempt to characterize the distribution of strengths of cotton fibers [8], and has since found a plethora of diverse applications as reviewed in [9]. Our approach will be reminiscent (yet distinct) from that of Fisher and Tippett, and will in fact closely follow the RG-inspired approach we used for deriving the Lévy stable distributions, albeit with the *cumulative* distribution function replacing the role of the *characteristic* function—for reasons that will shortly become clear. Note that other works in the literature also use an RG approach to study this problem, but in a rather different way (e.g. [3, 10–15]). While here the derivation only relies on the fact that taking the maximum through different procedures should lead to the same result (in the spirit of RG approaches), these works use a 'traditional' renormalization group approach; more specifically, in conventional RG one considers the flow of a variable, a number of variables, or in our context, an entire distribution (which is often referred to in the literature as 'functional RG' for this reason), upon a coarse-graining step. Having established such RG-flow, one can explore both the location as well as properties of the fixed points of the mapping (which in our case correspond to the possible limiting distributions), test their stability, and importantly, consider the flow to each of the fixed points. This often allows one to find the 'basin of attraction' of each of the fixed point—which for our problem determines which universality class a particular function belongs to. As mentioned below, such RG approaches can typically also provide us with the finite-size corrections to the fixed point, corresponding to 'truncating' the RG flow before it reached the fixed point. In contrast, the RG-inspired approach here, at least in its current form, can efficiently constrain the forms of the limiting distributions (the fixed point) but it is unclear whether it can also capture the finite-size corrections to it or provide us with more information akin to that provided by the RG-flow of the alternative approaches.

Extreme values. We will be interested in the maximum (or minimum) of a *large* number of variables. By nature, this (rare) random event is an outlier—the largest or smallest over many trials (assumed here to be independent). Indeed, the results are often applied to problems where the extreme events matter—what should be the height of a dam? What is the chance of observing an earth-quake or tsunami of a given magnitude? For these reasons insurance companies are likely to be interested in this topic.

To begin, we define:

$$X_n \equiv \max(x_1, x_2, ..., x_n), \tag{38}$$

where $x_1, ..., x_n$ are again i.i.d. variables. Since we have:

$$Prob(X_n < x) = Prob(x_1 < x)Prob(x_2 < x)...Prob(x_n < x) = C^n(x), \tag{39}$$

it is natural to work with the cumulative distribution when dealing with extreme value statistics, akin to the role which the characteristic function played in the previous section. Clearly, it is easy to convert the question of the *minimum* of n variables to one related to the maximum, if we define $\tilde{p}(x) = p(-x)$.

Before proceeding to the general analysis, which will yield three distinct universality classes (corresponding to the Gumbel, Weibull and Fréchet distributions), we will exemplify the behavior of each class on a particular example.

5.1. Example I: the Gumbel distribution

Consider the distribution:

$$p(x) = \mathrm{e}^{-x}. \tag{40}$$

Its cumulative is:

$$C(x) = 1 - \mathrm{e}^{-x}. \tag{41}$$

The cumulative distribution for the maximum of n variables is therefore:

$$G(x) = (1 - \mathrm{e}^{-x})^n \approx \mathrm{e}^{-n\mathrm{e}^{-x}} = \mathrm{e}^{-\mathrm{e}^{-(x-x_0)}}, \tag{42}$$

with $x_0 \equiv \log(n)$.

This is an example of the *Gumbel distribution*. The general form of its cumulative is:

$$G(x) = \mathrm{e}^{-\mathrm{e}^{-(ax+b)}}. \tag{43}$$

Taking the derivative of equation (42) to find the probability distribution for the maximum, we find:

$$p_n(x) = \mathrm{e}^{-\mathrm{e}^{-(x-x_0)}}\mathrm{e}^{-(x-x_0)}, \tag{44}$$

(where the n dependence enters only via x_0). Denoting $l \equiv \mathrm{e}^{-(x-x_0)}$, we have:

$$p(x) = \mathrm{e}^{-l}l, \tag{45}$$

and taking the derivative with respect to l we find that the distribution is peaked at $x = \log(n)$. It is easy to see that its width is of order unity. We can now revisit the approximation we made in equation (42), and check its validity.

Rewriting $(1 - x/n)^n = \mathrm{e}^{n\log(1-x/n)}$ and Taylor expanding the exponent to second order, we find that the approximation

$$(1 - x/n)^n \approx \mathrm{e}^{-x}, \tag{46}$$

is valid under the condition $x \ll \sqrt{n}$. In our case, this implies:

$$\mathrm{e}^{-x}n \ll \sqrt{n}. \tag{47}$$

J. Stat. Mech. (2020) 013214

At the peak of the distribution ($x = \log(n)$), we have $e^{-x}n = 1$, and the approximation is clearly valid there for $n \gg 1$. From equation (47) we see that the approximation we used would break down when we take x to sufficiently smaller than $\log(n)$. Defining $x = \log(n) - \delta x$, we see that the value of δx for which the approximation fails obeys $e^{-\delta x} = O(\sqrt{n})$, hence $\delta x = O(\log(\sqrt{n}))$. Since as we saw earlier the width of the distribution is of order unity, this implies that for large n the Gumbel distribution would approximate the exact solution well, failing only sufficiently far in the (inner) tail where the probability distribution is vanishingly small. However, a note of caution is in place: the logarithmic dependence we found signals a very slow convergence to the limiting form. This is also true in the case where the distribution $p(x)$ is Gaussian, as was already noted in Fisher and Tippett's original work [8]. For this reason the aforementioned more traditional RG approaches are very useful, as they are generally apt at going beyond the fixed point behavior only (here, the limiting distribution) and also capture the finite-size corrections to it [13, 14].

Finally, it is also interesting to explore the *tail* of the distribution of the maximum, where the above approximations fail. This is studied in [16] using a large-deviation theoretic approach.

5.2. Example II: the Weibull distribution

Consider the *minimum* of the same distribution we had in the previous example. The same logic would give us that:

$$Prob[min(x_1, ..x_n) > \xi] = Prob[x_1 > \xi]Prob[x_2 > \xi]..Prob[x_n > \xi] = e^{-\xi n}. \tag{48}$$

This is an example of the Weibull distribution, which occurs when the variable is bounded (e.g: in this case the variable is never negative).

As we shall see below, the general case, for the case of a maximum of n variables with distribution bounded by x^*, would be:

$$G(x) = \begin{cases} e^{-a(x^*-x)^{1/\alpha}}, & x \leqslant x^* \\ 0, & x > x^*. \end{cases} \tag{49}$$

In this case, the behavior of the original distribution $p(x)$ near the cutoff x^* is important, and determines the exponent α.

5.3. Example III: the Fréchet distribution

The final example belongs to the third possible universality class, corresponding to variables with a power-law tail.

If at large x we have:

$$p(x) = \frac{A_+}{(x-B)^{1+\mu}}. \tag{50}$$

Then the cumulative distribution is:

$$C(x) = 1 - \frac{A_+}{\mu(x-B)^{\mu}}. \tag{51}$$

J. Stat. Mech. (2020) 013214

Therefore taking it to a large power n we find:

$$C^n(x) \approx \mathrm{e}^{-\frac{A_+ n}{\mu(x-B)^{\mu}}}. \tag{52}$$

Upon appropriately scaling the variable, we find that:

$$G(x) = \mathrm{e}^{-a\left(\frac{x-b}{n^{1/\mu}}\right)^{-\mu}}, \tag{53}$$

(where a and b do not depend on n).

Importantly, we see that in this case the width of the distribution increases with n as a power-law $n^{1/\mu}$—for $\mu \leqslant 1$, this is precisely the same scaling we derived for the *sum* of n variables drawn from this heavy-tailed distribution! This elucidates why in the scenario $\mu \leqslant 1$ (corresponding to figure 1(c)) we obtained *Lévy flights*, where the sum was dominated by rare events no matter how large n was. This is related to the so-called 'single big jump principle', which has been recently shown to pertain to a broader class of scenarios in physics, extending the results for i.i.d. variables (see [17, 18] and references therein), as well as applications in finance [19].

We shall now show that these 3 cases can be derived in a general framework, using a similar approach to the one we used earlier.

5.4. General form for extreme value distributions

We will now find all possible limiting distributions, following similar logic to the RG-inspired approach used to find the form of the characteristic functions in the GCLT. By itself, our analysis will not reveal the 'basin of attraction' of each universality class, nor will we find the precise scaling of the coefficients a_n and b_n. These require work beyond the basic RG-inspired calculation presented here.

As before, let us assume that there exists some scaling coefficients a_n, b_n such that when we define:

$$\xi_n \equiv \frac{X_n - b_n}{a_n}, \tag{54}$$

the following limit exists:

$$\lim_{n\to\infty} Prob(\xi_n = \xi) = g(\xi). \tag{55}$$

(note that this limit is not unique: we can always shift and rescale by a constant). This would imply that $p(X_n) \approx a_n^{-1} g\left(\frac{X_n-b_n}{a_n}\right)$ and the cumulative is given by: $G\left(\frac{X_n-b_n}{a_n}\right)$. By the same logic we used before, we know that $G^m\left(\frac{X_n-b_n}{a_n}\right)$ depends only on the quantity $N = n \cdot m$. Therefore we have:

$$\frac{\partial}{\partial n}\left(G^{N/n}\left(\frac{X_n - b_n}{a_n}\right)\right) = 0. \tag{56}$$

(Note that here X_n is the random variable we are interested in: hence while derivatives of the *coefficients* a_n, b_n appear, a derivative of X_n is not defined and does not appear).

J. Stat. Mech. (2020) 013214

From which we find:

$$-\frac{N}{n^2}\log G + \frac{N}{n}\frac{G'}{G}\left[-\frac{\partial}{\partial n}\left(\frac{b_n}{a_n}\right) - \frac{X_n}{a_n^2}\frac{\partial a_n}{\partial n}\right] = 0. \tag{57}$$

Upon defining a new random variable $\tilde{x} \equiv \frac{X_n - b_n}{a_n}$ (as in equation (54)), the equation can be rewritten as:

$$[\log(-\log G(\tilde{x}))]' = -\left[\frac{n}{a_n}\frac{\partial b_n}{\partial n} + \frac{n}{a_n}\frac{\partial a_n}{\partial n}\tilde{x}\right]^{-1}. \tag{58}$$

In order for the RHS to have a sensible limit for large n, we would like to have:

$$\lim_{n\to\infty}\frac{n}{a_n}\frac{\partial a_n}{\partial n} = \alpha, \tag{59}$$

with α constant. Similarly, we have:

$$\lim_{n\to\infty}\frac{n}{a_n}\frac{\partial b_n}{\partial n} = \beta, \tag{60}$$

with β constant.

We shall shortly show that the value of α will dictate which of the three universality classes we will converge to.

Fréchet distribution:

If $\alpha > 0$, we find that to leading order (with the same subtle interpretation as in the case of the Lévy stable distribution above, see box 'A word of caution'):

$$a_n \propto n^{\alpha}. \tag{61}$$

Next, requiring that $\frac{n}{a_n}\frac{\partial b_n}{\partial n}$ should be constant implies that $b_n \propto a_n$. This corresponds to a *shift* in the scaled variable, and therefore we can set $b_n = 0$ without loss of generality.

Solving for $G(\tilde{x})$ gives the Fréchet distribution:

$$G(x) = \mathrm{e}^{-ax^{-1/\alpha}}. \tag{62}$$

Comparing this form with equation (53), we recognize that $1/\alpha = \mu$.

Weibull distribution:

Similarly, when $\alpha < 0$ we find that to leading order:

$$a_n \propto n^{-|\alpha|}. \tag{63}$$

Solving for G gives us the Weibull Distribution:

$$G(x) = \mathrm{e}^{a(\beta/|\alpha| - x)^{1/|\alpha|}}, \tag{64}$$

with a some constant.We see that for this class, the cumulative distribution equals precisely 1 at $x = \beta/|\alpha|$, implying that the distribution vanishes for x larger than this value (while it is manifestly non-zero for smaller x). Note that this threshold is arbitrary: we can always add a term proportional to a_n to the scaling coefficients b_n, and

J. Stat. Mech. (2020) 013214

shift the limiting distribution as desired. Indeed, according to equations (63) and (60) we have $b_n \approx \hat{x} + \beta\alpha n^{-1/|\alpha|} = \hat{x} + \hat{s}a_n$, with $\hat{x}$ and $\hat{s}$ constants. We conclude that for large n the coefficients b_n are approximately *constant*, and that the coefficients a_n monotonically *decrease* with n. Since for large n the cumulative of the original distribution is well approximated by $G\left(\frac{X_n - b_n}{a_n}\right)$, and since the support of the limiting distribution is bounded from above by $\beta/|\alpha|$, we conclude that the original distribution is also bounded from above by $\lim_{n\to\infty} a_n\beta/|\alpha| + b_n = \hat{x}$. The above relation implies that we are essentially 'zooming into' a particular region of the probability distribution as n increases. The distribution of the maximum becomes *narrower* as n increases, and is focused near its upper bound $\hat{x}$.

As mentioned before, the coefficient α is determined by the behavior of the original probability distribution near the cutoff x^* (which we showed above must equal $\hat{x}$). In the particular example discussed earlier $p(x)$ approached a non-zero constant near x^*, hence we found $|\alpha| = 1$. It is straightforward to generalize this to the case where $p(x)$ vanishes near the cutoff as a power-law $(x - x^*)^c$, finding that $1/|\alpha| = c + 1$.

Gumbel distribution:

Finally, consider the case $\alpha = 0$. Given that $\frac{n}{a_n}\frac{\partial b_n}{\partial n}$ is approximately constant, we obtain the Gumbel distribution:

$$G(x) = \mathrm{e}^{-\mathrm{e}^{-(ax+b)}}. \tag{65}$$

In the two previous cases where $\alpha \neq 0$, we found that the leading order behavior of the coefficients a_n, b_n was pinned-down by the RG-inspired approach. This is not the case when the RHS of equation (59) vanishes. In this case, unfortunately, the leading order of the scaling coefficients a_n is non-universal, and therefore cannot be determined from the RG-inspired approach alone. According to equation (60) the same holds for the scaling coefficients b_n. For $\alpha = 0$ both scaling coefficients a_n and b_n must be determined from the tail of $p(x)$. An interesting extended discussion can be found in [16], where the Gaussian case is analyzed. It can be shown that for the Gaussian distribution a particular (but non-unique) choice of scaling coefficients that leads to convergence to a Gumbel distribution is [8, 16]:

$$a_n = 1/b_n; b_n = \sqrt{2\log(n) - \log(4\pi\log(n))}. \tag{66}$$

We can now revisit equation (58), and plug in these explicit expressions for a_n and b_n. A straightforward calculation shows that:

$$\frac{n}{a_n}\frac{\partial b_n}{\partial n} = 1 - \frac{1}{2\log n}, \tag{67}$$

hence for large n it is approximately 1. In contrast:

$$\frac{n}{a_n}\frac{\partial a_n}{\partial n} = \frac{1 - \frac{1}{2\log n}}{2\log(n) - \log(4\pi\log(n))}, \tag{68}$$

therefore the second term of equation (58) indeed vanishes for large n—albeit very slowly due to the logarithmic dependence!

J. Stat. Mech. (2020) 013214

Another word of caution. Throughout the work, the logic of the RG-inspired approach *assumes* we have some scaling coefficients a_n, b_n such that the limit of equation (55) exists, and draws the (rather strong) constraints on the limiting distributions and scaling coefficients from this condition. But it is not *a priori* clear that such a scaling is at all possible! Indeed, consider a probability distribution with support on (e, ∞), and the following cumulative distribution:

$$C(x) = 1 - 1/\log(x). \tag{69}$$

The cumulative of the sum of n i.i.d variables drawn from this distribution will be:

$$C^n(x) = (1 - 1/\log(x))^n \approx \mathrm{e}^{-n/\log(x)} = \mathrm{e}^{-1/\log(x^{1/n})} \tag{70}$$

where the approximation is valid for $1/\log(x) \ll \sqrt{n}$. Clearly, linear scaling of the form of equation (54) would not be able to get rid of the n dependence in equation (70)—therefore convergence to one of the three universality classes does not occur in this scenario.

Another way of seeing this relies on the evaluation of a simple limit, the result of which tells us which (if any) universality class a given function belongs to (this and other useful relations are succintly summarized in [16]). Consider:

$$c = \log_2 \left[\lim_{\epsilon \to 0} \frac{C^{-1}(1-\epsilon) - C^{-1}(1-2\epsilon)}{C^{-1}(1-2\epsilon) - C^{-1}(1-4\epsilon)} \right]. \tag{71}$$

If the limit exists, a positive c implies convergence to Fréchet, negative to Weibull and 0 to Gumbel. In the above example, the limit diverges since $C^{-1}(1-\epsilon) = \mathrm{e}^{1/\epsilon}$ —hence the limiting distribution does not exist for *any* linear scaling, as we saw above.

6. Summary

The generalized central limit theorem and the extreme value distributions are often referred to as tales of tails—primarily dealing with distributions that are 'heavy-tailed', leading to the breakdown of the CLT. We began by exploring sums of (i.i.d.) random variables. We used a renormalization-group-inspired approach to find all possible limiting (stable) distributions of the sums, leading us to a generalization of the CLT to heavy-tailed distributions. Finally, we used a similar approach to study the similarly universal behavior of the *maximum* of a large number of (i.i.d) variables, in which case the cumulative distribution played the part previously taken by the characteristic function. In both cases the self-similarity of the resulting sum or maximum led to a simple ODE governing the limiting distributions and elucidating their universal properties. In the future, it would be interesting to see if this approach can be extended to functions of multiple variables [20], as well as to the case of *correlated* rather than i.i.d random variables, which although relevant to numerous applications has only recently began to be more systematically explored [21, 22].

J. Stat. Mech. (2020) 013214

Acknowledgments

I thank Ori Hirschberg for numerous useful comments and for a critical reading of this manuscript. I thank the students of AM 203 ('Introduction to disordered systems and stochastic processes') at Harvard University where this material was first tested, as well as the participants of the Acre Summer School on Stochastic Processes with Applications to Physics and Biophysics held in September 2017, and in particular Eli Barkai. I thank Ethan Levien, Farshid Jafarpour, Pierpaolo Vivo and the anonymous referees for useful comments on the manuscript.

Appendix. Code for running sum

In order to generate the data shown in figure 1(a), the following MATLAB code is used:

```
tmp = randn(N,1);
x = cumsum(tmp);
```

For figure 1(b), to generate a running sum of variables drawn from the Cauchy distribution, the first line is replaced with:

```
tmp = tan(pi*(rand(N,1)-1/2));
```

Finally, for figure 1(c) the same line is replaced with:

```
t = rand(N,1);
b = 1/mu;
tmp = t.^(-b);
```

where we used $\mu = 1/2$ for the figure.

References

[1] Feller W 2008 *An Introduction to Probability Theory and its Applications* vol 2 (New York: Wiley)
[2] Indyk P 2000 *Proc. 41st Annual Symp. on Foundations of Computer Science* (IEEE) pp 189–97
[3] Sethna J 2006 *Statistical Mechanics: Entropy, Order Parameters, and Complexity* vol 14 (Oxford: Oxford University Press)
[4] Jona-Lasinio G 2001 *Phys. Rep.* **352** 439
[5] Calvo I, Cuchí J C, Esteve J G and Falceto F 2010 *J. Stat. Phys.* **141** 409
[6] Klafter J and Sokolov I M 2011 *First Steps in Random Walks: from Tools to Applications* (Oxford: Oxford University Press)
[7] Taleb N N 2007 *The Black Swan: the Impact of the Highly Improbable* vol **2** (New York: Random house)
[8] Fisher R A and Tippett L H C 1928 *Mathematical Proceedings of the Cambridge Philosophical Society* vol 24 (Cambridge: Cambridge University Press) pp 180–90
[9] Fortin J-Y and Clusel M 2015 *J. Phys. A: Math. Theor.* **48** 183001
[10] Bertin E and Györgyi G 2010 *J. Stat. Mech.* P08022
[11] Calvo I, Cuchí J C, Esteve J G and Falceto F 2012 *Phys. Rev.* E **86** 041109
[12] Manzato C, Shekhawat A, Nukala P K, Alava M J, Sethna J P and Zapperi S 2012 *Phys. Rev. Lett.* **108** 065504
[13] Györgyi G, Moloney N, Ozogány K and Rácz Z 2008 *Phys. Rev. Lett.* **100** 210601
[14] Györgyi G, Moloney N, Ozogány K, Rácz Z and Droz M 2010 *Phys. Rev.* E **81** 041135
[15] Bazant M Z 2000 Largest cluster in subcritical percolation *Phys. Rev.* E **62** 1660
[16] Vivo P 2015 *Eur. J. Phys.* **36** 055037
[17] Vezzani A, Barkai E and Burioni R 2019 *Phys. Rev.* E **100** 012108
[18] Wang W, Vezzani A, Burioni R and Barkai E 2019 *Phys. Rev. Research* **1** 033172
[19] Filiasi M, Livan G, Marsili M, Peressi M, Vesselli E and Zarinelli E 2014 *J. Stat. Mech.* P09030
[20] Teuerle M, Żebrowski P and Magdziarz M 2012 *J. Phys. A: Math. Theor.* **45** 385002
[21] Schehr G and Le Doussal P 2010 *J. Stat. Mech.* P01009
[22] Majumdar S N, Pal A and Schehr G 2019 *Phys. Rep.* (https://doi.org/10.1016/j.physrep.2019.10.005)