

# Condensation for random variables conditioned by the value of their sum

Claude Godrèche

► **To cite this version:**

Claude Godrèche. Condensation for random variables conditioned by the value of their sum. Journal of Statistical Mechanics: Theory and Experiment, IOP Publishing, 2019, 2019 (6), pp.063207. 10.1088/1742-5468/ab1c4f. cea-03043347

**HAL Id: cea-03043347**

**<https://hal-cea.archives-ouvertes.fr/cea-03043347>**

Submitted on 7 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Condensation for random variables conditioned by the value of their sum

Claude Godrèche

Institut de Physique Théorique, Université Paris-Saclay, CEA and CNRS,  
91191 Gif-sur-Yvette, France

**Abstract.** We revisit the problem of condensation for independent, identically distributed random variables with a power-law tail, conditioned by the value of their sum. For large values of the sum, and for a large number of summands, a condensation transition occurs where the largest summand accommodates the excess difference between the value of the sum and its mean. This simple scenario of condensation underlies a number of studies in statistical physics, such as, e.g., in random allocation and urn models, random maps, zero-range processes and mass transport models. Much of the effort here is devoted to presenting the subject in simple terms, reproducing known results and adding some new ones. In particular we address the question of the quantitative comparison between asymptotic estimates and exact finite-size results. Simply stated, one would like to know how accurate are the asymptotic estimates of the observables of interest, compared to their exact finite-size counterparts, to the extent that they are known. This comparison, illustrated on the particular example of a distribution with Lévy index equal to  $3/2$ , demonstrates the role of the contributions of the dip and large deviation regimes. Except for the last section devoted to a brief review of extremal statistics, the presentation is self-contained and uses simple analytical methods.

## 1. Introduction

A question underlying a number of studies in statistical physics or in probability theory is the following. Let  $X_1, \dots, X_n$  be  $n$  independent, identically distributed (iid) positive random variables with finite mean. Assume that  $n$  is large and that the sum of these random variables is conditioned to take a fixed value, which can be smaller, equal to or larger than its mean. The question is to know how the (positive or negative) difference  $\Delta$  between the fixed value of the sum and its mean is distributed amongst the summands  $X_i$ , once a dependency between them has been introduced by the conditioning.

The answer to this question can be informally summarised as follows. If the common density of the random variables  $X_i$  is exponential, then, after conditioning, each of the summands takes a bit of the difference  $\Delta$ , whether negative or positive. The system is said to be in a ‘fluid phase’. If this density is subexponential (power law, stretched exponential), the same holds when the difference  $\Delta$  is negative. However, when it is positive (i.e., in excess) and large, in contrast to the exponential case, in general *only one* of the summands, the ‘condensate’, bears this excess. The remaining  $n - 1$  summands, which form the so-called ‘critical background’, are essentially unconstrained. This means that the dependency between the summands  $X_i$  introduced by the conditioning goes asymptotically in the condensate. One then speaks of a ‘condensation transition’. When  $\Delta = 0$  the system is again essentially made of a critical background.

This phenomenon can be illustrated by considering a random walk whose steps are the summands  $X_i$ , and which is conditioned to end at a given position at time  $n$ . Figure 1 depicts six histories of such a random walk with a power-law distribution of steps with tail index  $\theta = 3/2$ , conditioned to end at four times its mean,  $4n\langle X \rangle$ , at time  $n$ . For each trajectory one can observe the occurrence of a ‘big jump’ whose magnitude fluctuates around  $\Delta = 3n\langle X \rangle$ . In other words a large deviation of the sum is typically realised by a single big jump. The latter, i.e., the greatest summand, is the condensate referred to above. After removing this condensate the resulting histories are essentially unconstrained. In figure 1 one may note the presence of an history (in green) made of two big jumps. The role of such trajectories will be discussed in section 6 and later sections.

The analytical formulation of this question is as follows. The summands  $X_i$  are, from now on, except at the end of this paper, continuous random variables. Their common density is denoted by  $f_X(x)$ , with mean  $\langle X \rangle \equiv c_1$  ( $c_1$  is the first cumulant). Denoting by  $y$  the value taken by their sum,  $S_n = \sum_i X_i$ , the joint density of the  $X_i$  and of  $S_n$  is

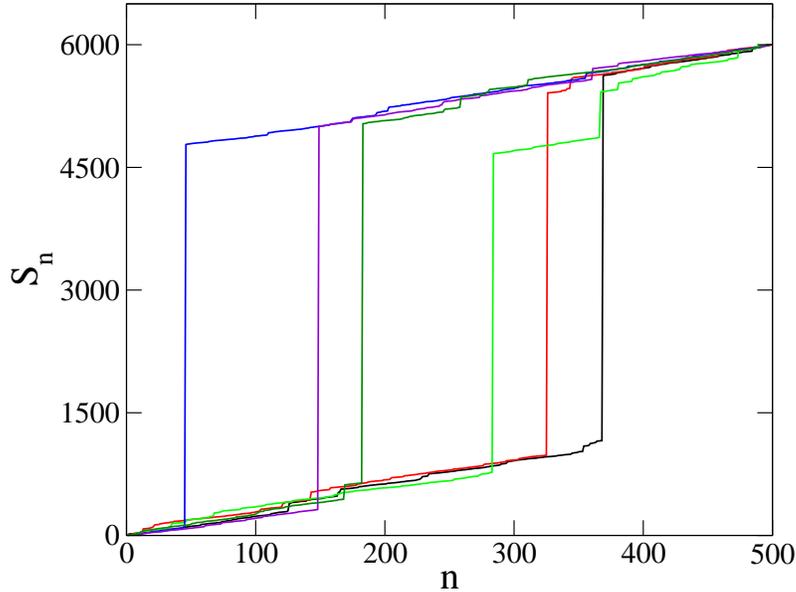
$$f(x_1, \dots, x_n, y) = f_X(x_1) \dots f_X(x_n) \delta\left(\sum_{i=1}^n x_i - y\right).$$

Summing upon all variables but  $y$  yields the density of  $S_n$ ,

$$f_n(y) = \frac{d}{dy} \text{Prob}(S_n < y) = \int dx_1 \dots dx_n f_X(x_1) \dots f_X(x_n) \delta\left(\sum_{i=1}^n x_i - y\right).$$

The joint conditional density of the random variables  $X_1, \dots, X_n$  under the condition  $S_n = y$ , denoted for short by  $f(x_1, \dots, x_n|y)$ , therefore reads

$$f(x_1, \dots, x_n|y) = \frac{f_X(x_1) \dots f_X(x_n) \delta(\sum_i x_i - y)}{f_n(y)},$$



**Figure 1.** Six trajectories of a random walk  $S_n = \sum_{i=1}^n X_i$  made of steps with power-law distribution  $f_X(x) = 3x^{-5/2}/2$  ( $x > 1$ ), for which  $\langle X \rangle = 3$ . The random walk is conditioned to end at position  $4\langle S_n \rangle = 6000$  at time  $n = 500$ . For each trajectory one can observe the occurrence of a ‘big jump’ whose magnitude fluctuates around  $\Delta = 3\langle S_n \rangle = 4500$ . Note however that the green history is made of two big jumps (see section 6 for a discussion of this point).

the presence of the denominator ensuring the normalisation.

We shall mainly be interested in the marginal conditional distribution of one of the  $X_i$ , denoted for short by  $f(x|y)$ , obtained from the previous expression by summing upon all  $X_i$  but one, to give

$$f(x|y) = f_X(x) \frac{f_{n-1}(y-x)}{f_n(y)}, \quad (1.1)$$

which can be interpreted as the “dressed” distribution of one of the  $X_i$  as opposed to the ‘bare’ distribution  $f_X(x)$ . The associated conditional average is thus

$$\langle X | S_n = y \rangle = \int_0^y dx f(x|y) = \frac{y}{n} \equiv \rho. \quad (1.2)$$

The difference  $\Delta$  between the value of the sum  $S_n$  and its mean  $\langle S_n \rangle = nc_1$  can be therefore simply expressed in terms of the difference between the conditional and unconditional averages

$$\Delta = y - nc_1 = n(\rho - c_1) \equiv n[\langle X | S_n = y \rangle - \langle X \rangle].$$

Looking again at figure 1, the marginal  $f(x|y)$  can be operationally seen as the limiting distribution of the summands (i.e., the step lengths of the random walk) for a large number of trajectories. Since the largest summand, the condensate, appears to be clearly separated from the other ones, the marginal  $f(x|y)$  is expected to have a hump shape in a neighbourhood of  $\Delta$ , representing the fluctuations of the condensate.

There are numerous studies related to this subject, dealing with urn models [1, 2, 3, 4], zero-range processes [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 4], mass-transport models [18, 19, 20], random allocation or random tree problems [21], to quote but a few. Large deviations for random walks with sub-exponential increments are considered in [22].

In the present work we revisit this very question with two more specific aims in mind. Firstly, we shall devote special care to the analysis of the distribution of the sum,  $f_n(y)$ , and of the marginal distribution of the summands,  $f(x|y)$ , in the various regimes of interest, with emphasis on the role of rare events. Secondly, for a particular example of power-law distribution of the summands, we shall confront the asymptotic predictions obtained for a large but finite number of summands to their exact counterparts. This gives a hint of the accuracy of the predictions of asymptotic analysis for more general distributions where exact finite-size expressions are not available.

In what follows we focus on the case where the density  $f_X(x)$  of the random variables  $X_i$  has a power-law tail,

$$f_X(x) \underset{x \rightarrow \infty}{\approx} \frac{c}{x^{1+\theta}}, \quad (1.3)$$

with  $\theta > 1$  in order to have a finite mean  $c_1$ . We shall however begin, in section 2, by the analysis of the simpler situation where  $f_X(x)$  is exponential, for which condensation does not occur. We shall then proceed by analysing the general case of a power-law distribution (1.3). As can be seen on the expression (1.1), the knowledge of the distribution of the sum,  $f_n(y)$ , allows to infer the marginal distribution  $f(x|y)$ . The detailed analysis of  $f_n(y)$  in the different regimes is therefore the building block for the study of the marginal  $f(x|y)$  (section 4). This analysis will be done in Laplace space, using the preparatory material contained in section 3. The results thus obtained are then applied, in section 5, to the special instance of the distribution (5.1) with power-law exponent  $\theta = 3/2$ , where exact expressions at finite  $n$  can be derived, in order to illustrate and validate the asymptotic analysis made in the general case of section 4. Section 6 is devoted to the derivation of the marginal distribution  $f(x|y)$  in the various regimes, both for a generic power-law distribution (1.3) and for the special case (5.1). The question of the unicity of the condensate and the statistics of extremes are reviewed in sections 7 and 8. The case of discrete random variables is summarised in Appendix A.

The present study builds upon previous works, especially [18, 19, 12], and consists, to a large extent, of an update of [19], with some effort devoted to giving a self-contained presentation, using simple analytical methods. It has no pretension to being exhaustive on all aspects of the field. In particular, reviewing the vast mathematical literature on sums of iid subexponential random variables and on the distribution of such random variables conditioned by a large value of their sum is beyond the scope of this work. The mathematical references most relevant to the present work are [8, 13, 15, 16, 17, 21], mentioned above. Let us finally mention [23], devoted to finite-size effects in zero-range condensation as manifested for example in the current overshoot, which shares some common features with the present work.

## 2. Exponentially distributed iid random variables

We start with the simple case of the exponential distribution

$$f_X(x) = \frac{e^{-x/c_1}}{c_1},$$

for which the distribution of the sum  $f_n(y)$  and the marginal  $f(x|y)$  are known exactly. First, the sum,  $S_n$ , has a gamma distribution

$$f_n(y) = \frac{y^{n-1} e^{-y/c_1}}{c_1^n \Gamma(n)}, \quad (2.1)$$

which is the inverse Laplace transform (with  $\text{Re } s > -1/c_1$ ) of

$$\hat{f}_n(s) = (\hat{f}_X(s))^n = \frac{1}{(1 + sc_1)^n},$$

as can be checked by inspection. Therefore the marginal distribution  $f(x|y)$  (1.1) is inferred from the exact expression (2.1) to give

$$f(x|y) = (n-1) \frac{(y-x)^{n-2}}{y^{n-1}}. \quad (2.2)$$

It does not depend on  $c_1$  and is monotonically decreasing with  $x$ , which is a manifestation of the absence of condensation. The conditional average  $\langle X|S_n = y \rangle$  (1.2) computed from (2.2) is equal to  $\rho$ , as it should.

Setting  $y = n\rho$  in (2.2) and letting  $n \rightarrow \infty$ , with  $\rho$  and  $x$  fixed yields the asymptotic estimate‡

$$f(x|y) \approx \frac{n}{y} (1 - x/y)^n \approx \frac{e^{-x/\rho}}{\rho}. \quad (2.3)$$

This estimate holds irrespectively of whether  $\rho$  is smaller or larger than  $\langle X \rangle = c_1$ . In other words, the system adjusts itself in such a way that the conditional distribution  $f(x|y)$  is still given by the ‘bare’ distribution,  $f_X(x)$ , with only a change of the parameter from  $c_1$  to  $\rho \leq c_1$ .

We now turn to the large deviation estimate of  $f_n(y)$ . We set, as above,  $y = n\rho$  in the expression (2.1) of  $f_n(y)$  and take the limit  $n \rightarrow \infty$ . This yields

$$f_n(y) \approx \frac{e^{n(1-\rho/c_1 + \ln \rho/c_1)}}{\sqrt{2\pi n} \rho}, \quad (2.4)$$

which reproduces the exact distribution (2.1) up to the replacement of  $\Gamma(n)$  by its Stirling approximation. With exponential accuracy we can write

$$f_n(y) \sim e^{-nI(\rho)},$$

where the large deviation function,

$$-I(\rho) = 1 - \frac{\rho}{c_1} + \ln \frac{\rho}{c_1}, \quad (2.5)$$

is defined for any value of the density  $\rho$  and is minimal and vanishes at  $\rho = c_1$ .

Using (2.5) yields an accurate estimate of  $f(x|y)$  for all values of  $x$ . In particular, (2.3) is recovered in the same limit as above, setting  $y = n\rho$  and letting  $n \rightarrow \infty$ , for  $\rho$  and  $x$  fixed.

‡ The symbol  $\approx$  stands for asymptotic equivalence. The symbol  $\sim$  stands either for ‘of the order of’, or for ‘with exponential accuracy’.

Anticipating on the sequel (compare to (6.3)), the rightmost expression in (2.3) can be recast as

$$f(x|y) \approx \frac{e^{-s_\rho x} f_X(x)}{\int_0^\infty dx e^{-s_\rho x} f_X(x)}, \quad (2.6)$$

where

$$s_\rho = \frac{1}{\rho} - \frac{1}{c_1}$$

can be positive, negative or zero. The denominator in (2.6) ensures normalisation. If we use (2.6) to compute the density  $\rho$  by (1.2), we find a relation between  $s_\rho$  and  $\rho$ ,

$$\rho \approx \frac{\langle X e^{-s_\rho X} \rangle}{\langle e^{-s_\rho X} \rangle}. \quad (2.7)$$

As shown later, (2.7) is the saddle-point equation for the inverse Laplace representation of  $f_n(y)$ .

To conclude, there is no condensation in the present case. The system is always in a fluid phase where, irrespectively of its sign, the difference  $\Delta$  is evenly distributed over all summands.

### 3. Laplace space and singularities

In what follows the asymptotic analysis of the distribution  $f_n(y)$  of the sum  $S_n$  is performed in Laplace space. The Laplace transform of  $f_n(y)$  with respect to  $y$  is

$$\hat{f}_n(s) = (\hat{f}_X(s))^n,$$

where  $\hat{f}_X(s) \equiv \langle e^{-sX} \rangle$ , hence, by inversion,

$$f_n(y) = \int_C \frac{ds}{2\pi i} e^{sy} (\hat{f}_X(s))^n, \quad (3.1)$$

where  $C$  is a Bromwich contour located on the right of the origin. The analysis of the distribution of the sum  $S_n$  therefore relies upon the analysis of the singularities of  $\hat{f}_X(s)$  in the complex  $s$ -plane. For the power-law distribution (1.3) the Laplace transform  $\hat{f}_X(s)$  has a cut extending along the negative real axis. When  $n$  is large  $(\hat{f}_X(s))^n$  is dominated by  $\hat{f}_X(s) \approx 1$ , i.e.,  $s$  small. The analytical structure of the Laplace transform  $\hat{f}_X(s)$  in the vicinity of the origin will therefore play a crucial role in the analysis of the distribution of  $S_n$ .

For a density  $f_X(x)$  with a power-law tail (1.3) the expansion of  $\hat{f}_X(s)$  for  $s \rightarrow 0$ , can be decomposed into a regular and a singular part

$$\hat{f}_{\text{reg}}(s) = 1 - sc_1 + \frac{s^2}{2} \langle X^2 \rangle + \dots, \quad (3.2)$$

$$\hat{f}_{\text{sing}}(s) = as^\theta + \dots, \quad (3.3)$$

where the parameter  $a$  is related to the tail parameter  $c$  by [24]

$$a = \Gamma(-\theta) c. \quad (3.4)$$

The parameter  $a$  is negative if  $0 < \theta < 1$ , positive if  $1 < \theta < 2$ , and so on. For instance,  $\Gamma(-1/2) = -2\sqrt{\pi}$ ,  $\Gamma(-3/2) = 4\sqrt{\pi}/3$ ,  $\Gamma(-5/2) = -8\sqrt{\pi}/15$ . The number of non-zero moments in the expansion of the regular part depends on the value of  $\theta$ . For  $1 < \theta < 2$  the first moment is defined, for  $2 < \theta < 3$  the second moment is also defined, and so

on. The expansion of the generating function of cumulants  $K(s) = \ln \hat{f}_X(s)$  follows from (3.2) and (3.3)

$$K(s) = -sc_1 + \frac{s^2}{2}c_2 - \dots + as^\theta + \dots, \quad (3.5)$$

where

$$c_2 \equiv \text{Var } X,$$

denotes the second cumulant. The first dots stand for higher-order regular terms ( $s^3, \dots$ ) and the second dots stand for higher-order singular terms ( $s^{\theta+1}, \dots$ ).

#### 4. Sum of iid positive random variables with a power-law tail

We now focus on the case where the density  $f_X(x)$  has a power-law tail (1.3) with exponent  $\theta$ . We will investigate successively the bulk of the distribution of  $S_n$  (generalised central limit theorem), then its left and right tails.

##### 4.1. Generalised central limit theorem

*Reminder.* We start with a reminder of well-known results on the generalised central limit theorem. By completeness we consider also the case where  $\theta < 1$ , though it is not relevant for the present study since the first moment  $\langle X \rangle = c_1$  is infinite.

The generalised central limit theorem [25] states that, for iid random variables with density (1.3), there exists two positive sequences  $a_n$  and  $b_n$  such that, when  $n \rightarrow \infty$ , the centered and scaled sum

$$U_n = \frac{S_n - b_n}{a_n}$$

converges (in distribution) to a stable law with index  $\alpha$ , where

$$\alpha = \begin{cases} \theta & \text{if } \theta < 2, \\ 2 & \text{if } \theta > 2, \end{cases} \quad (4.1)$$

and asymmetry parameter  $\beta = 1$ . Indeed, in the general case of a distribution  $f_X(x)$  with right and left power-law tails  $c_\pm/|x|^{1+\theta}$  ( $x \rightarrow \pm\infty$ ), the asymmetry parameter  $\beta$  is, by definition, the ratio  $(c_+ - c_-)/(c_+ + c_-)$ . In the present case of positive random variables the parameter  $c_- = 0$ , and  $\beta$  is thus equal to unity. We denote  $c_+$  by  $c$ , as in (1.3). If  $0 < \alpha < 2$ , this stable law also depends on the tail parameter  $c$ . If  $\alpha = 2$  the stable law is a Gaussian, the expression of which neither contains the asymmetry parameter  $\beta$  nor the tail parameter  $c$ .

The scale parameter  $a_n$  is equal to  $n^{1/\alpha}$ , where  $\alpha$  is given by (4.1), the centering parameter  $b_n$  is equal to  $nc_1$  when the mean is finite ( $\theta > 1$ ), and to zero otherwise ( $0 < \theta < 1$ ). Thus, for  $\theta > 2$  ( $\alpha = 2$ ), the usual central limit theorem is recovered,

$$\text{Prob}(u_1 \leq U_n \leq u_2) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi c_2}} \int_{u_1}^{u_2} du e^{-u^2/2c_2}, \quad (4.2)$$

while for  $0 < \theta < 2$  ( $\alpha = \theta$ ), the generalised central limit theorem reads

$$\text{Prob}(u_1 \leq U_n \leq u_2) \xrightarrow{n \rightarrow \infty} \int_{u_1}^{u_2} du L_{\alpha,c}(u),$$

where  $L_{\alpha,c}(u)$  is the density of the stable law of index  $\alpha$ , asymmetry parameter  $\beta = 1$  and tail parameter  $c$ . To summarise, the (generalised) central limit theorem gives the universal behaviour of the distribution of the sum  $S_n$  in the bulk, namely

$$f_n(y) \approx \frac{1}{n^{1/2}} G\left(\frac{y - nc_1}{n^{1/2}}\right), \quad (4.3)$$

if  $\theta > 2$  ( $\alpha = 2$ ), where  $G(u)$  is the Gaussian defined in (4.2),

$$f_n(y) \approx \frac{1}{n^{1/\theta}} L_{\theta,c}\left(\frac{y - nc_1}{n^{1/\theta}}\right), \quad (4.4)$$

if  $1 < \theta < 2$  ( $\alpha = \theta$ ), and

$$f_n(y) \approx \frac{1}{n^{1/\theta}} L_{\theta,c}\left(\frac{y}{n^{1/\theta}}\right), \quad (4.5)$$

if  $0 < \theta < 1$  ( $\alpha = \theta$ ).

*Examples.* For instance, for  $\alpha = 1/2$ , this distribution, the so-called Lévy law of index  $1/2$ , is explicit and reads

$$L_{1/2,c}(u) = \frac{c e^{-\pi c^2/u}}{u^{3/2}}, \quad (u > 0), \quad (4.6)$$

$$\hat{L}_{1/2,c}(s) = e^{-2c\sqrt{\pi s}}. \quad (4.7)$$

Another example, analysed in detail later, is the stable law with index  $\alpha = 3/2$ , which is explicitly given in terms of the Airy function (see (5.5)). More generally the Laplace transform of any stable law with index  $0 < \alpha < 2$  ( $\alpha \neq 1$ ) and asymmetry parameter  $\beta = 1$  reads

$$\hat{L}_{\alpha,c}(s) = e^{as^\alpha}, \quad (4.8)$$

where the parameter  $a$  is defined in (3.4). Thus in direct space

$$L_{\alpha,c}(u) = \int_C \frac{ds}{2\pi i} e^{su+as^\alpha}, \quad (4.9)$$

where  $C$  is a Bromwich contour located on the right of the origin. For  $0 < \alpha < 1$  the density of the stable law is only defined for  $u > 0$ , while for  $1 < \alpha < 2$  the support of the density is the whole real axis, implying that its Laplace transform is bilateral.

*Short proof of the generalised central limit theorem.* We start with the case  $1 < \theta < 2$ . The generating function of cumulants  $K(s)$  is, for small  $s$ , keeping the leading terms,

$$K(s) = \ln \hat{f}_X(s) \approx -sc_1 + as^\theta, \quad (4.10)$$

so, in this regime, the estimate of (3.1) is

$$f_n(y) \approx \int_C \frac{ds}{2\pi i} e^{s(y-nc_1)+nas^\theta}. \quad (4.11)$$

Setting

$$y - nc_1 = u n^{1/\theta}, \quad s = t n^{-1/\theta}, \quad (4.12)$$

yields (4.4), using (4.9). The regime considered here thus corresponds to  $\rho \rightarrow c_1$ . We proceed likewise for  $\theta > 2$ . Keeping the leading terms in the expansion of  $K(s)$ , we obtain

$$f_n(y) \approx \int_C \frac{ds}{2\pi i} e^{s(y-nc_1)+nc_2s^2/2}. \quad (4.13)$$

We now set

$$y - nc_1 = u\sqrt{n}, \quad s = t/\sqrt{n}, \quad (4.14)$$

which leads to the usual central limit theorem (4.3). The third case (4.5) can be proven likewise.

*Asymptotic behaviours of stable laws.* In both cases (i.e., if either  $0 < \alpha < 1$  or  $1 < \alpha < 2$ )  $L_{\alpha,c}(u)$  has the same right tail (1.3) as the initial distribution  $f_X(x)$ ,

$$L_{\alpha,c}(u) \underset{u \rightarrow \infty}{\approx} \frac{c}{u^{1+\alpha}}, \quad (0 < \alpha < 2), \quad (4.15)$$

as can be seen by linearising the integrand of (4.9) with respect to  $s^\alpha$ , and folding the contour around the negative real axis (see for details in section 4.3 where the same reasoning is used).

The asymptotic behaviour of the stable law on the left can be obtained by the saddle-point method. We have

$$L_{\alpha,c}(u) \underset{u \rightarrow 0}{\approx} \frac{B}{u^\nu} e^{-A/u^\mu}, \quad (0 < \alpha < 1), \quad (4.16)$$

$$L_{\alpha,c}(u) \underset{u \rightarrow -\infty}{\approx} B|u|^\nu e^{-A|u|^\mu}, \quad (1 < \alpha < 2), \quad (4.17)$$

with exponents

$$\mu = \frac{\alpha}{|1-\alpha|}, \quad \nu = \frac{2-\alpha}{2|1-\alpha|},$$

and where the two positive constants  $A$  and  $B$  read

$$A = \frac{|1-\alpha|}{\alpha} (\alpha|a|)^{1/(1-\alpha)}, \quad B = \frac{(\alpha|a|)^{1/(2(1-\alpha))}}{\sqrt{2\pi|1-\alpha|}}.$$

For example, if  $\alpha = 1/2$ , the asymptotic estimate (4.16) reproduces identically the whole law (4.6). For  $\alpha = 3/2$  we obtain, using (3.4),

$$L_{3/2,c}(u) \underset{u \rightarrow -\infty}{\approx} \frac{\sqrt{|u|}}{2\pi c} e^{-|u|^3/(12\pi c^2)}, \quad (4.18)$$

a result related to (5.6) below.

*Away from the bulk.* The generalised central limit theorem does not predict the behaviour of the distribution of the sum  $S_n$  in the tails. We now investigate the behaviour of  $f_n(y)$  away from the bulk, that is, when the difference  $|\Delta| = |y - nc_1|$  is extensive, i.e., of order  $n$ , (while in the regimes (4.12 and (4.14) it was subextensive), first to the left ( $y < nc_1$ ), then to the right ( $y > nc_1$ ), restricting the study to the case  $\theta > 1$ , such that  $c_1$  is finite.

#### 4.2. Left tail: large deviations

The left tail of  $f_n(y)$  corresponds to those rare events where  $\rho < c_1$ , hence  $\Delta$  large and negative. In this regime, which is far away from the regime of validity of the generalised central limit theorem, the large deviation estimate of the density  $f_n(y)$  is non universal and depends on the details of the distribution  $f_X(x)$ . We first present the general framework for the computation of the large deviation function  $I(\rho)$  (4.22), valid for any  $\theta > 1$ . There is no explicit expression of this function in general for distribution of the type (1.3). We shall later find an explicit expression of this large deviation function for the distribution (5.1) with tail index  $\theta = 3/2$ , valid in all regimes (see section 5). For the time being, we will content ourselves with the expressions (4.26) and (4.27) of the large deviation function in the scaling regime where  $\rho$  is close to  $c_1$ , for a general distribution (1.3). Equation (4.26) restores the generalised central limit theorem in the regime (4.17). Equation (4.27) restores the usual central limit theorem.

*General framework.* Let us come back on (3.1) that we recast as

$$f_n(y) = \int_C \frac{ds}{2\pi i} e^{sy+nK(s)} = \int_C \frac{ds}{2\pi i} e^{-nI(\rho,s)}, \quad (4.19)$$

with

$$-I(\rho, s) = s\rho + K(s). \quad (4.20)$$

If  $n$  is large it is natural to perform a saddle-point analysis of (4.19). The saddle-point equation reads

$$\left. \frac{dI(\rho, s)}{ds} \right|_{s_\rho} = 0,$$

that is to say<sup>§</sup>

$$-K'(s_\rho) = \frac{\langle X e^{-s_\rho X} \rangle}{\langle e^{-s_\rho X} \rangle} = \rho. \quad (4.21)$$

The position of the saddle point  $s_\rho$  on the real axis depends on the value of  $\rho = y/n$ . This saddle point only exists if  $\rho < c_1$ . Indeed, if  $\rho = c_1$ , the saddle point  $s_\rho = 0$  hits the head of the cut of  $K(s)$  (see (3.5)), hence the saddle-point equation (4.21) cannot be satisfied beyond  $\rho = c_1$ . Defining the large deviation function as

$$-I(\rho) \equiv -I(\rho, s_\rho) = s_\rho \rho + K(s_\rho), \quad (4.22)$$

we finally obtain

$$f_n(y) \underset{n \rightarrow \infty}{\approx} \frac{e^{-nI(\rho)}}{\sqrt{2\pi n K''(s_\rho)}}, \quad (\rho < c_1), \quad (4.23)$$

with

$$K''(s) = \frac{\langle X^2 e^{-sX} \rangle}{\langle e^{-sX} \rangle} - \left( \frac{\langle X e^{-sX} \rangle}{\langle e^{-sX} \rangle} \right)^2. \quad (4.24)$$

<sup>§</sup> The saddle-point equation (4.21) was anticipated in (2.7).

*Scaling regime.* Determining the large deviation function in the scaling region  $\rho \rightarrow c_1$  implies expanding the expressions above for  $s_\rho \rightarrow 0$ .

- We start with  $1 < \theta < 2$ . The saddle-point equation

$$-K'(s_\rho) \approx c_1 - a\theta s_\rho^{\theta-1} = \rho,$$

yields

$$s_\rho \approx \left( \frac{c_1 - \rho}{a\theta} \right)^{1/(\theta-1)}, \quad (4.25)$$

only defined if  $\rho < c_1$ . We thus find, using (4.10), the expression of the large deviation function in this regime,

$$I(\rho) \approx \frac{\theta - 1}{\theta(a\theta)^{1/(\theta-1)}} (c_1 - \rho)^{\theta/(\theta-1)}. \quad (4.26)$$

The right side of this equation can be identified with  $A|u|^\mu/n$  in (4.17). Actually, in this scaling regime, the full large deviation estimate (4.23) reduces to (4.4) with (4.17). The left tail (4.17) can indeed be seen as the large deviation estimate of  $L_{\theta,c}(u)$ .

The special case of  $\theta = 3/2$  is treated in detail in section 5. The universal part of the large deviation function (4.26) gives (5.7).

- For  $\theta > 2$  we have  $-K'(s_\rho) \approx c_1 - c_2 s_\rho = \rho$ , hence  $s_\rho \approx (c_1 - \rho)/c_2$ , yielding the quadratic form

$$I(\rho) \approx \frac{(c_1 - \rho)^2}{2c_2}. \quad (4.27)$$

Thus (4.23), with  $K''(s_\rho) \approx c_2$ , gives the central limit theorem (4.3) back.

*Remark.* The two equations (4.22) and (4.23) provide a parametric representation of  $I(\rho)$ , i.e., of  $K'(s_\rho)s_\rho - K(s_\rho)$  against  $-K'(s_\rho)$  (which is  $\rho$ ), which can be used for numerical purposes.

#### 4.3. Right tail: ‘deep in the condensed phase’

Again the regime considered here, where  $\rho > c_1$ , is different from that prevailing for the central limit theorem. Recall that, for any value of  $\theta > 1$ , using (3.5),

$$f_n(y) = \int_{\mathcal{C}} \frac{ds}{2\pi i} e^{s(y-nc_1)+\dots+nas^\theta+\dots}. \quad (4.28)$$

Now  $\Delta = n(\rho - c_1)$  is of order  $n$ , so  $s \sim 1/n$ , implying that  $ns^\theta \sim n^{1-\theta}$  is subextensive. Therefore the two terms  $s(y - nc_1)$  and  $nas^\theta$  are no longer balanced as in (4.11) and (4.12). The contour  $\mathcal{C}$  is deformed to encircle the real negative axis. The leading contribution to  $f_n(y)$  comes from linearising with respect to the leading singular term:

$$f_n(y) \approx na \int_{\mathcal{C}} \frac{ds}{2\pi i} e^{s(y-nc_1)} s^\theta.$$

Using the Hankel representation of the reciprocal Gamma function

$$\frac{1}{\Gamma(\theta)} = \int_{\mathcal{C}} \frac{ds}{2\pi i} e^s s^{-\theta},$$

we obtain

$$f_n(y) \approx \frac{na}{\Gamma(-\theta) (y - nc_1)^{1+\theta}}.$$

Finally, using (3.4), we have, for any value of  $\theta > 1$ , if  $\rho > c_1$ , thence for  $y - nc_1 \sim n$ ,

$$f_n(y) \approx \frac{nc}{(y - nc_1)^{1+\theta}}, \quad (4.29)$$

where  $c$  is the tail coefficient of  $f_X(x)$ . Similar considerations can be found in [12, 11].

This result matches with the asymptotic estimate (4.15) for  $y - nc_1 = n^{1/\theta}u$  ( $u$  large), if  $1 < \theta < 2$  (see (4.4)). This prediction holds further away in the tail, where the excess difference is extensive. Furthermore (4.29) also holds for  $\theta > 2$ . In other words, while at the scale  $y - nc_1 \sim n^{1/2}$  the tail is Gaussian, at the scale  $y - nc_1 \sim n$  it is given by (4.29). Equating (4.3) and (4.29) shows that the matching between the two behaviours occurs for

$$y - nc_1 \sim \sqrt{(\theta - 2)c_2} \sqrt{n \ln n}. \quad (4.30)$$

See [17] for related considerations.

As a last comment, let us remark that the contributions coming from the next terms  $(nas^\theta)^p$  in the expansion of  $e^{nas^\theta}$  in (4.28) are subleading by successive factors  $n^{-(p-1)(\theta-1)}$  with respect to the contribution of the first term  $nas^\theta$ . These subleading probabilities will be recovered otherwise in section 7.

### 5. The example of a distribution with power-law tail exponent $\theta = 3/2$

For the distribution

$$f_X(x) = \frac{2}{\sqrt{\pi}} \frac{e^{-1/x}}{x^{5/2}}, \quad (x \geq 0), \quad (5.1)$$

such that  $\langle X \rangle \equiv c_1 = 2$ , the exact distribution of the sum  $S_n$  is explicit and reads [19]

$$f_n(y) = \frac{ne^{-n^2/y}}{\sqrt{\pi}y^{(n+3)/2}} \left[ H_n \left( \frac{y+2n}{2\sqrt{y}} \right) - \sqrt{y}H_{n-1} \left( \frac{y+2n}{2\sqrt{y}} \right) \right], \quad (5.2)$$

where the  $H_n$  are Hermite polynomials. This exact result will provide an illustration of the statements made in the previous section as well as a benchmark for the asymptotic estimates given there. In Laplace space

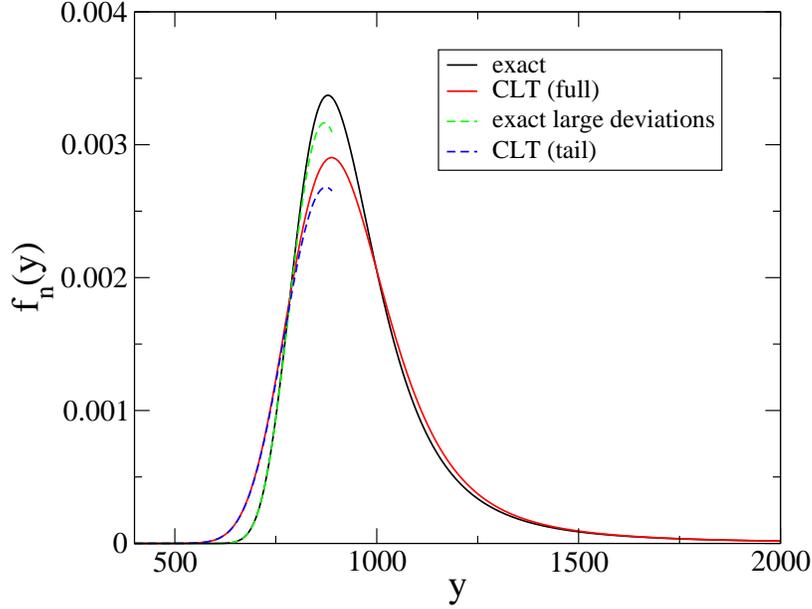
$$\hat{f}_X(s) = (1 + 2\sqrt{s})e^{-2\sqrt{s}},$$

as can be found by taking the derivative of (4.6) and (4.7) with respect to the tail parameter  $c$ . So, for small  $s$ ,

$$\hat{f}_X(s) \approx 1 - 2s + \frac{8}{3}s^{3/2}, \quad (5.3)$$

which is the beginning of the expansion  $1 - sc_1 + as^\theta + \dots$ , with  $c_1 = 2$  and  $a = 8/3$  obtained from (3.4) for  $c = 2/\sqrt{\pi}$  and  $\theta = 3/2$ . The generating function of cumulants is thus equal to

$$K(s) = \ln \hat{f}_X(s) \approx -2s + \frac{8}{3}s^{3/2}.$$



**Figure 2.** Density  $f_n(y)$  of the sum  $S_n$  of  $n = 500$  random variables with density (5.1) ( $\theta = 3/2$ ), against  $y$ . We compare the exact density (5.2), the prediction (5.4) of the generalised central limit theorem (CLT), the exact large deviation estimate (5.11) and the estimate (5.6) of this expression in the scaling regime  $\rho \lesssim c_1 = 2$ , which is also the left tail estimate of (5.4). (See the text for comments and the short summary below.)

*Central limit theorem.* The generalised central limit theorem states that the bulk (i.e., for  $\rho \approx 2$ ) of the distribution of the sum  $S_n$  is given by (4.4),

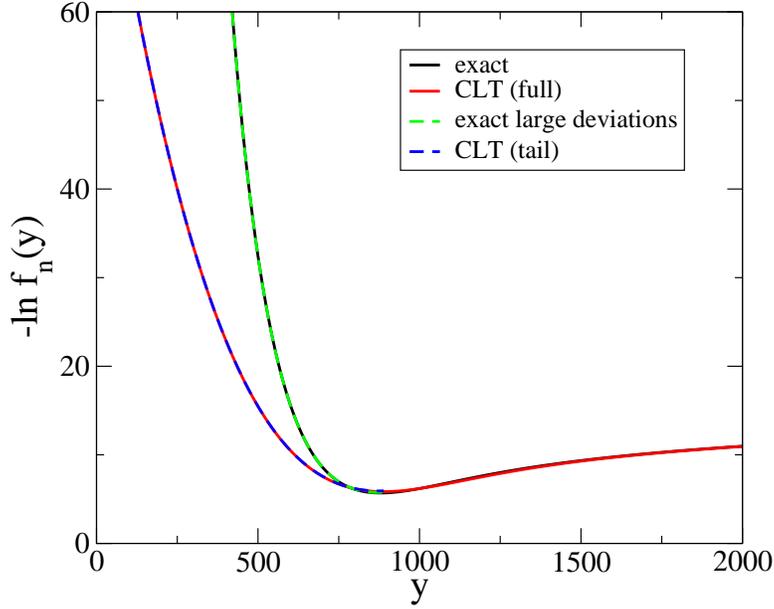
$$f_n(y) \approx \frac{1}{n^{2/3}} L_{3/2,c} \left( \frac{y - 2n}{n^{2/3}} \right), \quad (5.4)$$

where the stable law  $L_{3/2,c}(u)$  is explicitly known in terms of the Airy function and its derivative [26]. With  $c = 2/\sqrt{\pi}$  it reads

$$L_{3/2,c}(u) = -\frac{1}{2} \exp \left( \frac{u^3}{96} \right) \left[ \frac{u}{4} \text{Ai} \left( \frac{u^2}{16} \right) + \text{Ai}' \left( \frac{u^2}{16} \right) \right], \quad (5.5)$$

with Laplace transform

$$\hat{L}_{3/2,c}(s) = e^{\frac{8}{3}s^{3/2}}.$$



**Figure 3.** Same as figure 2 in linear-log. (See the text for comments.)

*Left tail of (5.4).* For  $u = (y - 2n)/n^{2/3}$  large negative, the behaviour of  $L_{3/2,c}(u)$  is given by (4.18), thus, in this regime, (5.4) yields

$$f_n(y) \approx \frac{\sqrt{2 - y/n}}{4\sqrt{n\pi}} e^{-n(2 - y/n)^3/48}. \quad (5.6)$$

Comparing (5.6) to the general expression (4.23) yields the large deviation function  $I(\rho)$  in the scaling regime  $\rho \lesssim 2$ ,

$$I(\rho) \approx \frac{1}{48}(2 - \rho)^3. \quad (5.7)$$

This expression, which is universal, is a particular form of (4.26), with  $\theta = 3/2$  and  $c_1 = 2$ .

*The large deviation function.* Following the scheme given in section 4.2 for the determination of the full large deviation function yields the saddle-point equation (4.21)

$$-K'(s_\rho) = \frac{2}{1 + 2\sqrt{s_\rho}} = \rho, \quad (5.8)$$

hence

$$\sqrt{s_\rho} = \frac{1}{\rho} - \frac{1}{2}, \quad (5.9)$$

confirming that the saddle point only exists for  $\rho < c_1 = 2$ . For  $\rho = c_1$ , the saddle-point value  $s_\rho$  vanishes. We thus find the expression of the large deviation function (as defined in (4.22)), which reads

$$-I(\rho) \equiv s_\rho \rho + K(s_\rho) = \frac{\rho}{4} - \frac{1}{\rho} + \ln \frac{2}{\rho}, \quad (\rho < 2), \quad (5.10)$$

and  $K''(s_\rho) = \rho^3/(2 - \rho)$ . Using (4.23), we finally obtain

$$f_n(y) \approx \frac{\sqrt{2-\rho}}{\sqrt{2\pi n \rho^3}} \exp \left[ n \left( \frac{\rho}{4} - \frac{1}{\rho} + \ln \frac{2}{\rho} \right) \right] \quad (5.11)$$

$$= \frac{n(2n)^n e^{-n^2/y+y/4} \sqrt{2-y/n}}{\sqrt{2\pi} y^{n+3/2}}. \quad (5.12)$$

Two remarks are in order. Firstly, for  $\rho \rightarrow c_1$ , i.e.,  $s_\rho \rightarrow 0$ ,  $K''(s_\rho) \rightarrow \infty$ . The reason is that, according to (4.24),  $K''(0) = \text{Var } X$ , which is infinite in the present case. Hence one does not expect good accuracy of this prediction when approaching  $c_1$ . Secondly, the expansion of (5.10) for  $\rho \lesssim 2$  yields (5.7) as it should. In this regime the large deviation estimate (5.11) takes the universal form (5.6).

*Right tail of  $f_n(y)$ .* When the difference  $\Delta = y - 2n$  is positive and extensive, the distribution of  $S_n$  is given by (4.29), with  $c = 2/\sqrt{\pi}$ , that is

$$f_n(y) \approx \frac{2n}{\sqrt{\pi}(y-2n)^{5/2}}. \quad (5.13)$$

*Remark: asymptotics of  $f_n(y)$  in the tails.* The results (5.11) (left tail) and (5.13) (right tail) can also be obtained by a direct asymptotic analysis of the exact expression (5.2), as we now show. In (5.2) the argument of the Hermite polynomial,

$$z = \frac{y+2n}{2\sqrt{y}},$$

defines a function  $z(y)$  which is minimum at  $y = 2n$ , where  $z = \sqrt{2n}$ . For  $y$  smaller or greater than  $2n$ ,  $z$  is always larger than  $\sqrt{2n}$ . We therefore need an asymptotic estimate of  $H_n(z)$  for  $z > \sqrt{2n}$ . This is obtained by a saddle-point analysis of the generating function of Hermite polynomials yielding (see Appendix B)

$$H_n(z) \approx e^{(z^2 - zV - n)/2} (z+V)^n \sqrt{(1+z/V)/2}, \quad V = \sqrt{z^2 - 2n}. \quad (5.14)$$

Using this estimate in (5.2), then setting  $y = n\rho$  with  $\rho < 2$ , and expanding for  $n \rightarrow \infty$  yields (5.11). Likewise setting  $y = n\rho$  with  $\rho > 2$ , then expanding for  $n \rightarrow \infty$ , yields (5.13).

*Numerical comparisons of exact predictions and asymptotic estimates.* In figures 2 and 3 we compare the analytical prediction (5.2) for the distribution of the sum of  $n = 500$  random variables with density (5.1) and tail index  $\theta = 3/2$ , with

- ★ the prediction (5.4) of the generalised central limit theorem,
- ★ the full large deviation estimate (5.11),

|| In the language of a quantum harmonic oscillator, this means that the region explored in the variable  $z$  when  $y$  varies from zero to infinity is the forbidden region where the Hermite polynomials do not oscillate.

★ and the estimate (5.6) of the latter in the scaling regime  $\rho \lesssim c_1 = 2$ ; equation (5.6) is equivalently the estimate for the left tail of the scaling form (5.4).

These figures illustrate the following facts:

- (i) The right tail of the exact expression (5.2) is in excellent numerical agreement with the right tail of the scaling form (5.4). (See the comments below (4.29).)
- (ii) The left tail of the exact expression (5.2) is in excellent numerical agreement with the large deviation estimate (5.11).
- (iii) The left tail of the central limit expression (5.4) is in excellent numerical agreement with its estimate (5.6).

Figure 4 depicts the centered and scaled exact result (5.2) for  $n = 125, 250, 500$ , together with the stable law (5.5), illustrating the slow convergence of the former to the latter.

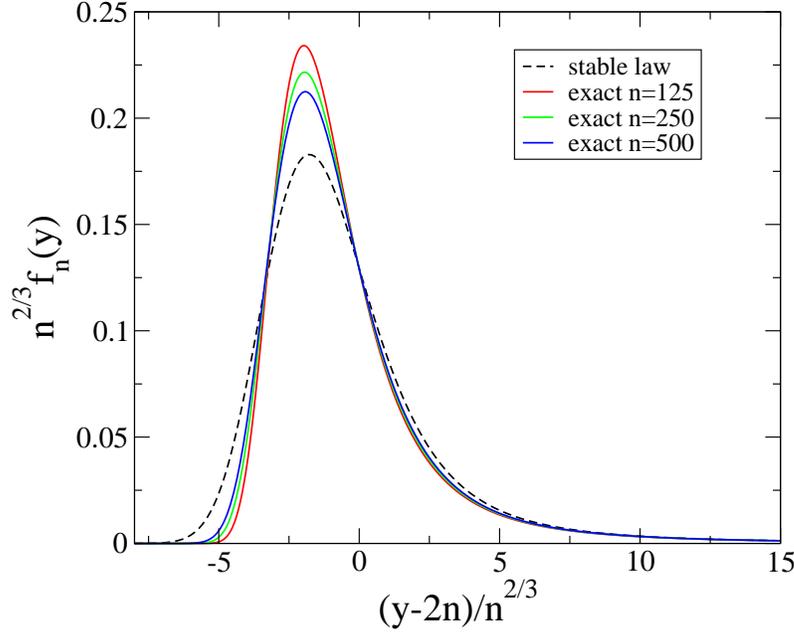
*A short summary.* The main equations obtained in this section and in section 4 can be summarised as follows,

	EXACT	(5.2)	–	–
	CLT (FULL)	(5.4)	(4.4)	(4.3)
	LD (FULL)	(5.11)	–	–
	DEEP	(5.13)	(4.29)	(4.29)
{	CLT (TAIL)	(5.6)	(4.23), (4.26)	(4.23), (4.27)
	LD (SCALING)			

These equations are identified by short names or acronyms in the left column (CLT: central limit theorem, LD: large deviation, DEEP: deep in the condensed phase). The second column refers to results concerning the distribution (5.1), the third column refers to results concerning the generic case (1.3), with  $1 < \theta < 2$ , and the rightmost column refers to the case  $\theta > 2$ .

In the generic case (1.3) no exact expression for the distribution  $f_n(y)$ , as in (5.2), is known. Neither is there in general an exact expression of the full large deviation estimate, as in (5.11). The (generalised) central limit theorem reproduces correctly the behaviour of the left tail of  $f_n(y)$  in the universal scaling region only, i.e., for  $y$  close to  $nc_1$ . For  $\Delta$  negative and extensive, only the full large deviation estimate is faithful, which, as said above, is not explicitly known in general. The right tail expression (4.29) is valid for any  $\theta > 1$ .

*Comparison with the case of a discrete distribution.* Finally, to complete this study, figure A1 depicts a comparison between the exact density  $f_n(y)$  (5.2) and its discrete counterpart, the partition function  $Z_{L,N}$  of the zero range process with hopping rate (A.15), where  $b = 5/2$ . The partition function is obtained recursively using (A.14). The curves are centered and scaled, in order to highlight the universality of the bulk in the continuum limit. The parameter  $r$  is the ratio of the tail parameters of the two functions, namely  $c = 2/\sqrt{\pi}$  for the first one and  $c = (b - 1)\Gamma(b) = 9\sqrt{\pi}/8$  for the second one (see (A.16) or (A.17)). The lower plot demonstrates the non universality of the large deviations in the left tail.



**Figure 4.** Comparison between the stable law (5.5) and the exact result (5.2) centered and scaled for  $n = 125, 250, 500$ .

## 6. Marginal conditional density and condensation

We are now in position to compute the marginal conditional distribution (1.1), repeated here for convenience,

$$f(x|y) = f_X(x) \frac{f_{n-1}(y-x)}{f_n(y)}, \quad (6.1)$$

where the density  $f_X(x)$  is given by (1.3). This conditional density is a function of  $x$ , while  $y$  plays the role of a parameter. We thus have to study separately  $f(x|y)$  for the different regimes of  $y/n = \rho$ . The study hereafter parallels that made in [19].

*Subcritical regime* ( $\rho < c_1 \Leftrightarrow \Delta < 0$ ). We start again from (3.1). Thus

$$f_{n-1}(y-x) = \int_C \frac{ds}{2\pi i} e^{sy+nK(s)-sx-K(s)}.$$

Let us assume that  $x$  is of order 1. So, at the saddle point, for  $n$  large, we have, within exponential accuracy (see section 4.2),

$$f_{n-1}(y-x) \sim e^{-nI(\rho)-s_\rho x-K(s_\rho)}, \quad (6.2)$$

where  $s_\rho$  satisfies the equation  $-K'(s_\rho) = \rho$ . This yields, for any  $\theta > 1$ , the handy expression

$$f(x|y) \underset{n \rightarrow \infty}{\approx} \frac{f_X(x)e^{-s_\rho x}}{\hat{f}_X(s_\rho)}, \quad (6.3)$$

which is well normalised and has its first moment equal to  $\rho$ . Its physical interpretation is appealing: there is ‘compression’ of the  $X_i$ , since each one of them bears a part of the negative difference  $\Delta$ . This accounts for the *fluid phase*. When  $x$  becomes large (6.2) and (6.3) are no longer correct. It is necessary to use the large deviation estimate (4.23) in order to obtain an accurate expression of the marginal density (6.1).

This study can be illustrated on the example of  $f_X(x)$  given by (5.1) ( $\theta = 3/2$ ). Equation (6.3) yields (using the accurate expression (5.9) for  $s_\rho$ )

$$f(x|y) \underset{n \rightarrow \infty}{\approx} \frac{\rho}{\sqrt{\pi}x^{5/2}} \exp\left(-\left[\frac{(2-\rho)x + 2\rho}{2\rho\sqrt{x}}\right]^2\right). \quad (6.4)$$

This expression is in excellent numerical agreement with the exact prediction for  $f(x|y)$  derived from (5.2) if  $x$  is of order 1, as soon as  $n$  is large enough. In contrast, the estimate obtained for  $f(x|y)$  using the scaling estimate (4.25) for  $s_\rho$  compares well to the true distribution only when  $\rho$  is not too far away from  $c_1$ . Finally, if  $x$  is no longer of order 1, the large deviation estimate (5.12) inserted into (6.1) provides an accurate estimate of the marginal distribution  $f(x|y)$ . Starting from this very expression, setting  $y = n\rho$  and letting  $n \rightarrow \infty$  restores (6.4), since  $x$  becomes  $\ll y$  in this limit.

**Critical regime** ( $\rho = c_1 \Leftrightarrow \Delta = 0$ ). Note that if  $\rho = c_1 = 2$ , then  $s_\rho = 0$  and both asymptotic estimates (6.3) and (6.4) reduce to  $f_X(x)$ . These estimates are obtained in the limit  $n \rightarrow \infty$  (in order for the saddle-point method to be valid). Therefore the reduction of  $f(x|y)$  to  $f_X(x)$  only holds in this limit. Otherwise there are finite-size corrections given by the expressions (6.5) and (6.6) below, where the estimate of  $f_n(y)$  in the bulk is used. For  $1 < \theta < 2$ ,

$$f(x|y) \approx f_X(x) \frac{L_{\theta,c}((c_1 - x)/n^{1/\theta})}{L_{\theta,c}(0)}, \quad (6.5)$$

and for  $\theta > 2$ ,

$$f(x|y) \approx f_X(x) \frac{G((c_1 - x)/n^{1/2})}{G(0)} = f_X(x) e^{-(x-c_1)^2/2nc_2}. \quad (6.6)$$

Again, if  $n \rightarrow \infty$ , one recovers the fact that  $f(x|y) \rightarrow f_X(x)$ . For  $x$  of order  $n$ , one should use the large deviation estimate (4.23) for  $f_{n-1}(y-x)$  (e.g. (5.11) for  $f_X(x)$  given by (5.1), with  $\theta = 3/2$ ).

**Supercritical regime** ( $\rho > c_1 \Leftrightarrow \Delta > 0$ ). In this regime  $f_n(y)$  is always given by its right-tail estimate (4.29)

$$f_n(y) \approx \frac{nc}{\Delta^{1+\theta}}. \quad (6.7)$$

The discussion therefore only focusses on  $f_{n-1}(y-x)$ , where  $x$  should be compared to  $\Delta$ , which is of order  $n$ . Beyond the obvious regime where  $x$  is of order unity, hence  $f(x|y) \approx f_X(x)$ , there are three other regimes to consider, corresponding respectively to the bulk, the right-tail and the large deviations of  $f_{n-1}(y-x)$ .

- (a) **Condensate.** If  $x \approx \Delta$  (that is  $\Delta - x \sim n^{1/\alpha}$ ), the ratio of  $f_X(x) \approx c/\Delta^{1+\theta}$  to  $f_n(y)$  given by (6.7) yields one piece of  $f(x|y)$

$$\frac{f_X(x)}{f_n(y)} \approx \frac{c/\Delta^{1/\theta}}{nc/\Delta^{1/\theta}} = \frac{1}{n}.$$

The other piece,  $f_{n-1}(y-x)$ , is given by its bulk since  $y-x \approx nc_1$ . Hence, if  $1 < \theta < 2$ ,

$$f(x|y)|_{\text{cond}} \approx \frac{1}{n} f_{n-1}(y-x) \approx \frac{1}{n} \frac{1}{n^{1/\theta}} L_{\theta,c} \left( \frac{\Delta-x}{n^{1/\theta}} \right), \quad (6.8)$$

and, if  $\theta > 2$ ,

$$f(x|y)|_{\text{cond}} \approx \frac{1}{n} f_{n-1}(y-x) \approx \frac{1}{n} \frac{1}{n^{1/2}} G \left( \frac{\Delta-x}{n^{1/2}} \right). \quad (6.9)$$

These expressions describe the bulk of the fluctuating condensate which manifests itself by a hump shape of the marginal  $f(x|y)$  for  $x \approx \Delta$  on figure 5. For any  $\theta > 1$  we have, from (6.8) or (6.9),

$$\int_{x \in \text{hump}} dx f(x|y)|_{\text{cond}} \approx \frac{1}{n}, \quad (6.10)$$

which demonstrates that the excess difference  $\Delta$  is borne by only one summand. (See also the discussion in section 7.)

- (b) **Dip.** The range of values of  $x$  such that  $x \gg 1$ ,  $\Delta - x \gg 1$ , interpolates between the critical part of  $f(x|y)$ , for  $x$  or order 1, and the condensate, for  $x$  close to  $\Delta$ . It corresponds to the *dip region* on figure 5. In this region,  $f_{n-1}(y-x)$  is given by its right tail (4.29) or (6.7). So, for any  $\theta > 1$ ,

$$f(x|y)|_{\text{dip}} \approx c \left[ \frac{\Delta}{x(\Delta-x)} \right]^{1+\theta} \approx \frac{f_X(x)f_X(\Delta-x)}{f_X(\Delta)}. \quad (6.11)$$

The interpretation of this result is that in the dip region typical configurations where one summand takes the value  $x$  are such that the remaining  $\Delta - x$  excess difference is borne by a single other summand. The dip region is therefore dominated by configurations where the excess difference is shared by *two* summands [12].

The weight of these configurations can be estimated as follows. Let  $\xi$  be some positive number less than  $1/2$ . Then

$$\int_{\xi\Delta}^{(1-\xi)\Delta} dx f(x|y)|_{\text{dip}} = \int_{\xi\Delta}^{(1-\xi)\Delta} dx \frac{f_X(x)f_X(\Delta-x)}{f_X(\Delta)} \sim \Delta^{-\theta} \sim n^{-\theta}. \quad (6.12)$$

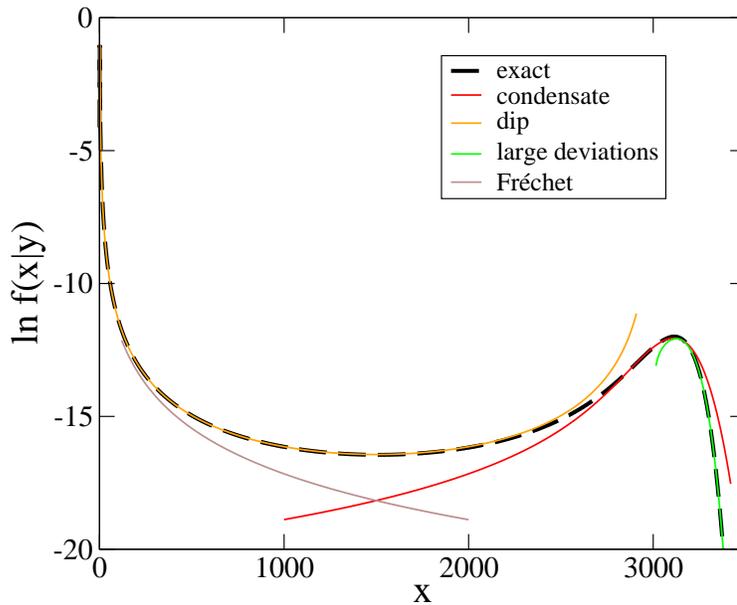
The relative weights of the dip and condensate regions is therefore of order  $n^{-(\theta-1)}$ , i.e., the weight of events where the condensate is broken in two pieces of order  $n$  is subleading with respect to events with a single big jump. This will be restated in section 7. The reduction factor  $n^{-(\theta-1)}$  is the same as that met in the discussion at the end of section 4.

An illustration of this phenomenon is given in figure 1. The overwhelming contribution to the statistics of trajectories comes from those exhibiting a single big jump of order  $n$ , approximately equal to  $\Delta$ . Some rare trajectories, as the green one, exhibit two big jumps instead of a single one, both of order  $n$ . These trajectories contribute to (6.11).

(c) **Large deviations.** Finally, if  $x > \Delta$ , one should use the large deviation estimate for  $f_{n-1}(y-x)$  (e.g. (5.11) for  $f_X(x)$  given by (5.1)).

In summary, the contribution of the condensate to the total weight is equal to  $1/n$ . The contribution of the dip region is subleading by a power-law factor. The contribution of the large deviations is exponentially subleading. The main contribution comes from the region where  $x$  is of order unity where  $f(x|y) \approx f_X(x)$ .

*Quantitative comparison.* Figure 5 summarises this study. It depicts the marginal distribution  $f(x|y)$ , with  $f_X(x)$  given by (5.1), for  $n = 500$ ,  $y = 4nc_1$ ,  $\Delta = y - nc_1 = 3000$  ( $c_1 = 2$ ). The curves named condensate, dip and large deviations correspond respectively to the cases (a), (b) and (c) above. The curve named Fréchet represents  $f^{(2)}(x)/n$  as defined in (8.1) and will be commented on in section 8.



**Figure 5.** Linear-log plot of the marginal distribution  $f(x|y)$  in the supercritical regime for  $f_X(x)$  given by (5.1) ( $\theta = 3/2$ ). Here  $n = 500$ ,  $\rho = y/n = 4c_1 = 8$ , hence  $\Delta \equiv y - nc_1 = 3000$ . The exact marginal distribution  $f(x|y)$  is reproduced by the union of three pieces, respectively the condensate, the dip and the large deviation contributions. The curve named Fréchet represents  $f^{(2)}(x|y)/n$  as defined in (8.1) (see section 8).

## 7. Unicity of the condensate

The analysis of the marginal distribution  $f(x|y)$  made in section 6 showed that the distribution  $f(x|y)$  has a hump shape for  $x \approx \Delta$ , the weight of which is equal to  $1/n$  according to (6.10). This means that the largest summand is the only one to ‘bear’ the excess difference  $\Delta$  and therefore that asymptotically the condensate is unique.

However, as discussed below (6.11), there exist configurations where the excess difference is shared by two summands (i.e., with now a leader and a subleader instead of a unique condensate) and whose weight is subleading by a factor of order  $n^{-(\theta-1)}$  with respect to configurations with a single big jump. Such configurations are those which dominate in the dip region.

We present hereafter another argument in favour of the unicity of the condensate which is independent of that recalled above, even if it is akin to it. The aim is to show that the event with a unique  $X_i$  bearing all the excess difference  $\Delta$  is much more likely than the event corresponding to two summands  $X_i$  sharing it. This issue has been previously discussed in [11] for discrete variables, in the context of the statics of the zero-range process. Uniqueness of the condensate has also been established rigorously in the discrete and continuous cases in [15] and [16] (see also [10]).

The probability associated to the event where  $X_1$  bears the excess difference is

$$\text{Prob}(\Delta < X_1 < \Delta + dx | S_n = y) = f(\Delta|y)dx, \quad (7.1)$$

with

$$f(\Delta|y) = f_X(\Delta) \frac{f_{n-1}(y - \Delta)}{f_n(y)}, \quad (7.2)$$

and where  $y - \Delta = nc_1$ . This probability has to be multiplied by a factor  $n$  since any of the  $X_i$  can be chosen to bear the excess difference.

The probability corresponding to the event where  $X_1$  and  $X_2$  are both large and share the excess difference  $\Delta$  reads

$$\text{Prob}(\Delta < X_1 + X_2 < \Delta + dx | S_n = y) = \left( \int_{\xi\Delta}^{(1-\xi)\Delta} dx' f(x', \Delta - x'|y) \right) dx, \quad (7.3)$$

with

$$f(x, \Delta - x|y) = f_X(x)f_X(\Delta - x) \frac{f_{n-2}(y - \Delta)}{f_n(y)}, \quad (7.4)$$

and where  $\xi$  is some positive number less than  $1/2$ , as in (6.12). The probability (7.3) has to be multiplied by the binomial coefficient  $\binom{n}{2}$  which counts the possible choices of two  $X_i$  amongst  $n$ . The ratio  $f_{n-1}(nc_1)/f_{n-2}(nc_1)$  is asymptotically equal to one, so remains to estimate

$$nf_X(\Delta) \sim n\Delta^{-1-\theta} \sim n^{-(\theta-1)-1}, \quad (7.5)$$

and

$$\binom{n}{2} \int_{\xi\Delta}^{(1-\xi)\Delta} dx f_X(x)f_X(\Delta - x) \sim n^2\Delta^{-1-2\theta} \sim n^{-2(\theta-1)-1}. \quad (7.6)$$

The ratio of these two estimates scales as  $n^{\theta-1} \gg 1$  as soon as  $\theta > 1$ , which is precisely the condition for the existence of a condensate. This result can be generalised to the case of  $p$  variables sharing the excess difference  $\Delta$ . We now have to estimate

$$\binom{n}{p} \int dx_1 \dots dx_p f(x_1) \dots f(x_p) \delta\left(\sum_i x_i - \Delta\right) \sim n^p \frac{\Delta^{p-1}}{\Delta^{p(1+\theta)}} \sim n^{-p(\theta-1)-1}.$$

Thus the ratio of (7.5) to the latter yields  $n^{(p-1)(\theta-1)}$ .

*Remarks.*

- (i) The factor  $n^{\theta-1}$  is precisely that found at the end of section 4 by a different line of reasoning.
- (ii) Performing the integral in (7.6) from 0 to  $\Delta$  would yield  $f_2(\Delta)$  which scales as  $\Delta^{-1-\theta}$  instead of  $\Delta^{-1-2\theta}$  as in (7.6). Multiplied by  $\binom{n}{2}$  this yields  $n^{1-\theta}$ , which dominates (7.5) by a factor  $n$ , as it should.
- (iii) As a last remark, let us note that the ratio of (7.2) to (7.4) gives

$$\frac{f(x, \Delta - x|y)}{f(\Delta|y)} \approx \frac{f_X(x)f_X(\Delta - x)}{f_X(\Delta)},$$

which is the expression (6.11) for  $f(x|y)|_{\text{dip}}$ .

## 8. Largest summands

Investigating the statistics of extremes for the problem at hand is a natural question since the condensate is the largest summand. A number of works have been devoted to this question [8, 13, 15, 20, 21]. The discussion hereafter concerns the case where  $f_X(x)$  has a power-law tail (1.3).

In [8] the greatest summand is proven to scale as  $n$  in the supercritical regime, as  $n^{1/\theta}$  in the critical regime and as  $\ln n$  in the subcritical regime. In [13] it is shown that if the largest summand is removed, the measure on the remaining summands converges to the product measure with density  $\rho = c_1$ , when the number of summands  $n$  is fixed and the value of the sum  $y$  increases to infinity. This means that the remaining background is critical, a feature which is apparent in figure 1, as already mentioned.

Let us denote the  $k$ -th largest summand by  $X^{(k)}$  ( $k = 1, \dots, n$ ). The densities of these ranked summands, denoted by  $f^{(k)}(x|y)$ , sum up to

$$\sum_{k=1}^n f^{(k)}(x|y) = n f(x|y).$$

The distribution of the largest summand  $X^{(1)}$  is investigated in [20, 15, 21]. The result is that, if  $y = n\rho$ ,  $\rho > c_1$ ,  $n \rightarrow \infty$ , the rescaled variable

$$Z_n = n^{-1/\alpha}(\Delta - X^{(1)})$$

converges to a stable law of index  $\alpha$ , with  $\alpha$  defined in (4.1) (i.e.,  $\alpha = \theta$  if  $\theta < 2$  or  $\alpha = 2$  if  $\theta > 2$ ). This means that, asymptotically, the density of  $X^{(1)}$  coincides, up to a factor  $n$ , with the estimates of the marginal density in the condensate region ( $\Delta - x \sim n^{1/\alpha}$ ), that is with (6.8) or (6.9) according to the value of  $\theta$ ,

$$f^{(1)}(x|y) \approx n f(x|y)|_{\text{cond}}, \quad (\Delta - x \sim n^{1/\alpha}).$$

This result conforms with the intuition that, in the condensate region, the only contribution to the marginal  $f(x|y)$  comes from the largest summand.

One can already guess from the statements made in [8, 13] and recalled above that the distribution of the second largest summand,  $X^{(2)}$ , should be asymptotically Fréchet, and that the subsequent ones,  $X^{(k)}$  ( $k \geq 2$ ), should be the order statistics of  $n-1$  iid random variables  $X_i$  with density  $f_X(x)$  (i.e., before conditioning), which can be summarised by saying that, in the supercritical regime, the dependency between the summands  $X_i$  introduced by the conditioning goes asymptotically in the big jump  $X^{(1)}$ . Reference [21] indeed states that the rescaled variables

$$W_n^{(k)} = n^{-1/\theta} X^{(k)}, \quad (k \geq 2),$$

have asymptotic densities

$$f_W^{(k)}(w) = \frac{c}{w^{1+\theta}} e^{-c/(\theta w^\theta)} \frac{[c/(\theta w^\theta)]^{k-2}}{(k-2)!},$$

independently of the value of  $y$ . Hence, for  $k \geq 2$ ,

$$f^{(k)}(x|y) \approx \frac{1}{n^{1/\theta}} f_W^{(k)}\left(\frac{x}{n^{1/\theta}}\right) = \frac{nc}{x^{1+\theta}} e^{-nc/(\theta x^\theta)} \frac{[nc/(\theta x^\theta)]^{k-2}}{(k-2)!}.$$

For instance the curve named Fréchet in figure 5 represents

$$\frac{1}{n} f^{(2)}(x|y) \approx \frac{2}{\sqrt{\pi} x^{5/2}} e^{-4n/(3\sqrt{\pi} x^{3/2})}. \quad (8.1)$$

Since  $X^{(1)}$  typically scales as  $n$ , while  $X^{(2)}, X^{(3)}, \dots$  typically scale as  $n^{1/\theta}$ , the condensate is increasingly separated from the background as  $n$  increases, leaving space to the dip region ( $x \gg 1, \Delta - x \gg 1$ ). We know from the analysis made in section 6 (see discussion following (6.11)) that this region is dominated by configurations where the excess difference is shared by two summands, namely  $X^{(1)}$  and  $X^{(2)}$ , so

$$f^{(1)}(x|y) + f^{(2)}(x|y) \approx n f(x|y)|_{\text{dip}}, \quad (x \gg 1, \Delta - x \gg 1), \quad (8.2)$$

and that the contributions of these events to  $n f(x|y)$  are of order  $n^{-(\theta-1)}$ . To the right of  $\Delta/2$  the predominant contribution to the sum on the right side of (8.2) comes from  $f^{(1)}(x|y)$ , to the left it comes from  $f^{(2)}(x|y)$ . In this respect it is worth noting that, right in the middle of the dip, i.e., for  $x = \Delta/2$ , the following relations hold, if  $1 < \theta < 2$ ,<sup>¶</sup>

$$\begin{aligned} f(x|y)|_{\text{dip}} &\approx 4^{1+\theta} \frac{c}{\Delta^{1+\theta}}, \\ f(x|y)|_{\text{cond}} &\approx \frac{1}{n} f^{(1)}(x|y) \approx \frac{1}{n} f^{(2)}(x|y) \\ &\approx f_X(x) \approx 2^{1+\theta} \frac{c}{\Delta^{1+\theta}}, \end{aligned}$$

where  $f(x|y)|_{\text{cond}}$  is continued outside its region of validity ( $\Delta - x \sim n^{1/\theta}$ ). The ratio between the two quantities on the left side of the equations is therefore a universal number, only depending on the tail exponent  $\theta$ . Up to adding a tail correction to  $f^{(1)}(x|y)$  the same results are equally valid for  $\theta > 2$ .

*Remark.* The random variable  $Z_n$  is scaled by  $n^{1/\alpha}$ , where  $\alpha$  is defined in (4.1), while the random variables  $W_n^{(k)}$  are scaled by  $n^{1/\theta}$ . In the first case  $\alpha$  saturates at  $\alpha = 2$ , in the second case  $\theta$  can take any value.<sup>+</sup>

## 9. Discussion

In this work we have revisited the statistics of iid random variables with a power-law distribution (1.3) conditioned by the value of their sum. For large values of the latter, a condensation transition occurs where the largest summand accommodates the excess difference between the value of the sum and its mean. This simple scenario of

<sup>¶</sup> The crossing of  $f(x|y)|_{\text{cond}}$  and  $f^{(2)}(x|y)/n$  at  $x = \Delta/2$  is visible on figure 5.

<sup>+</sup> Compare to theorem 19.34 in [21] where the distinction between these two exponents is not made.

condensation underlies a number of studies in statistical physics, usually formulated in terms of discrete random variables such as, e.g., in random allocation and urn models, or condensing zero-range processes at stationarity. The present study extends easily to other subexponential distributions of the summands.

Much of the effort here has been devoted to presenting the subject in simple terms, reproducing known results (especially from [19] and [12]) and adding some new ones. In particular the comparison between asymptotic estimates and their finite-size counterparts demonstrates the role of the contributions of the dip and large deviation regimes. The contribution of the dip region is of crucial importance for the analysis of the stationary dynamics of the condensate [12]. The conclusions given in [12] have been confirmed by rigorous mathematical studies [27, 28, 29, 30].

To close, let us mention several related topics or generalisations of interest in [31, 32, 33, 34].

### Acknowledgments

It is a pleasure to thank J M Luck for many interesting discussions, S N Majumdar for discussions on ref. [19] and an anonymous referee for suggestions to improve the text.

### Appendix A. Discrete formalism

All the questions investigated so far with continuous random variables have a transcription in the language of discrete random variables. The resulting framework is that used in the description of equilibrium urn models in statistical mechanics or in the analysis of the stationary state of zero range processes. We successively review these three facets of the subject. Table A1 summarises the correspondences between the discrete and continuum formalisms.

#### Appendix A.1. Discrete random variables conditioned by the value of their sum

Let  $N_1, N_2, \dots, N_L$  be iid positive discrete random variables with distribution

$$\pi_k = \text{Prob}(N_1 = k), \quad (\text{A.1})$$

and average

$$\langle N_1 \rangle = \sum_{k \geq 0} k \pi_k = \rho_c. \quad (\text{A.2})$$

The joint distribution of these random variables reads

$$\text{Prob}(N_1 = n_1, \dots, N_L = n_L) = \pi_{n_1} \dots \pi_{n_L}. \quad (\text{A.3})$$

Assume now that their sum, denoted by  $S_L$ , is conditioned to be equal to  $N$ . Then the joint distribution of  $\{N_1, N_2, \dots, N_L\}$  and  $S_L$  is

$$\text{Prob}(N_1 = n_1, \dots, N_L = n_L, S_L = N) = \pi_{n_1} \dots \pi_{n_L} \delta\left(\sum_i n_i, N\right). \quad (\text{A.4})$$

Summing this expression on  $n_1, \dots, n_L$  yields the distribution of  $S_L$ , or partition function  $\tilde{Z}_{L,N}$ ,

$$\begin{aligned} \tilde{Z}_{L,N} &\equiv \text{Prob}(S_L = N) \\ &= \sum_{\{n_i\}} \pi_{n_1} \dots \pi_{n_L} \delta\left(\sum_i n_i, N\right). \end{aligned} \quad (\text{A.5})$$

The conditional joint distribution of  $N_1, N_2, \dots, N_L$ , given  $S_L$ , is the ratio of (A.4) to (A.5), that is

$$\text{Prob}(N_1 = n_1, \dots, N_L = n_L | S_L = N) = \frac{1}{\tilde{Z}_{L,N}} \pi_{n_1} \dots \pi_{n_L} \delta\left(\sum_i n_i, N\right), \quad (\text{A.6})$$

from which the marginal conditional distribution of one of the  $N_i$  (taken conventionally to be  $N_1$ ), denoted by  $f_k$ , ensues by summation

$$f_k = \text{Prob}(N_1 = k | S_L = N) = \pi_k \frac{\tilde{Z}_{L-1, N-k}}{\tilde{Z}_{L,N}}. \quad (\text{A.7})$$

The conditional average is thus

$$\langle N_1 | S_L = N \rangle = \sum_{k \geq 0} k f_k = \frac{N}{L} = \rho, \quad (\text{A.8})$$

by definition of the density  $\rho$ . Summing (A.7) on  $k$  leads to a recursion relation on the  $\tilde{Z}_{L,N}$

$$\tilde{Z}_{L,N} = \sum_{k \geq 0} \pi_k \tilde{Z}_{L-1, N-k}. \quad (\text{A.9})$$

**Table A1.** Correspondences between the discrete formalism of Appendix A.1 and the continuum formalism used in the bulk of the paper.

Discrete r.v.	Continuous r.v.
$L$	$n$
$N_1, \dots, N_L$	$X_1, \dots, X_n$
$\text{Prob}(N_i = k) = \pi_k$	$f_X(x)$
$\langle N_1 \rangle = \rho_c$	$\langle X \rangle = c_1$
$S_L = N_1 + \dots + N_L$	$S_n = X_1 + \dots + X_n$
$N$	$y$
$\rho = N/L$	$\rho = y/n$
$\text{Prob}(S_L = N) = \tilde{Z}_{L,N}$	$f_n(y)$
$f_k = \text{Prob}(N_1 = k   S_L = N)$	$f(x y)$
$\Pi(z)$	$\hat{f}_X(s)$
$\tilde{Z}_{L,N} \sim e^{-L\mathcal{F}(\rho)}$	$f_n(y) \sim e^{-nI(\rho)}$

*Thermodynamic limit.* In the thermodynamic limit the large deviation function (or free energy) reads

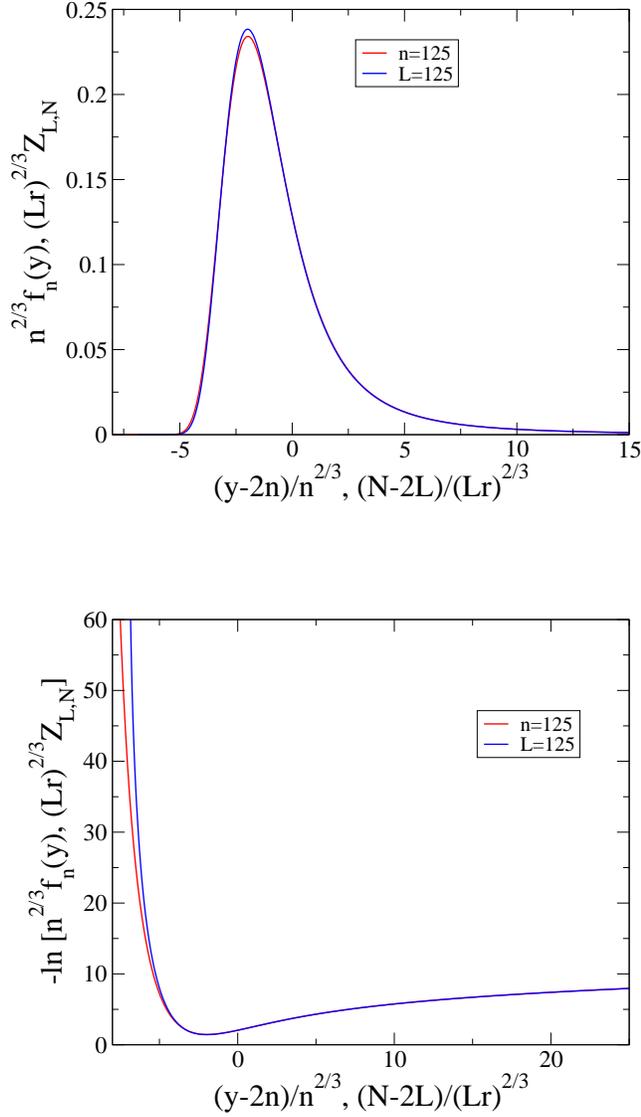
$$\mathcal{F} = - \lim_{L, N \rightarrow \infty} \frac{1}{L} \ln \tilde{Z}_{L,N},$$

i.e., with exponential accuracy,

$$\tilde{Z}_{L,N} = \text{Prob}(S_L = N) \sim e^{-L\mathcal{F}(\rho=N/L)}.$$

The large deviation function can be computed by the saddle-point method. Casting the integral representation of the Kronecker function

$$\delta(m, n) = \oint \frac{dz}{2\pi i z^{n+1}} z^m,$$



**Figure A1.** Comparison between the exact density  $f_n(y)$  (5.2), for  $n = 125$ , and the partition function  $Z_{L,N}$  of the ZRP with hopping rate (A.15) where  $b = 5/2$ , for  $L = 125$ , obtained recursively using (A.14). The curves are centered and scaled. The parameter  $r$  is the ratio of the tail parameters of the two functions (see text at the end of section 5). The lower figure is a linear-log plot of the upper one.

in (A.5) yields

$$\tilde{Z}_{L,N} = \text{Prob}(S_L = N) = \oint \frac{dz}{2\pi i z^{N+1}} \Pi(z)^L, \quad (\text{A.10})$$

where  $\Pi(z)$  is the generating function of the  $\pi_k$

$$\Pi(z) = \langle z^{N_1} \rangle = \sum_k z^k \pi_k.$$

The contour integral in (A.10) can be evaluated by the saddle-point method. The saddle-point equation is

$$\frac{z_\rho \Pi'(z_\rho)}{\Pi(z_\rho)} = \rho,$$

where the saddle-point value  $z_\rho$  depends on the density  $\rho$  through this equation. The discussion of this equation is analogous to that given in the continuum formalism.

### Appendix A.2. Equilibrium urn models

The framework described in the previous section is naturally realised by classical urn models, defined as follows. Consider a finite connected graph, made of  $L$  sites (or urns), on which  $N$  particles are distributed. The number of particles on site  $i$  is the random variable  $N_i$ , with  $S_L = \sum_{i=1}^L N_i = N$ . A configuration of the system is defined by the values  $\{n_1, \dots, n_L\}$ , taken by the random occupations  $N_1, \dots, N_L$ . The energy of such a configuration is the sum of the individual energies at each site,

$$E(\{n_i\}) = \sum_{i=1}^L E(n_i).$$

The associated unnormalised Boltzmann weight attached to site  $i$  is

$$p_{n_i} = e^{-\beta E(n_i)}.$$

The probability of the configuration  $\{n_i\}$  is therefore given by the product form

$$\text{Prob}(N_1 = n_1, \dots, N_L = n_L | S_L = N) = \frac{1}{Z_{L,N}} p_{n_1} \cdots p_{n_L} \delta\left(\sum_i n_i, N\right), \quad (\text{A.11})$$

where

$$Z_{L,N} = \sum_{\{n_i\}} p_{n_1} \cdots p_{n_L} \delta\left(\sum_i n_i, N\right), \quad (\text{A.12})$$

is the canonical partition function of this statistical mechanical system. The single-site occupation probability is

$$f_k = \text{Prob}(N_1 = k | S_L = N) = p_k \frac{Z_{L-1, N-k}}{Z_{L,N}}, \quad (\text{A.13})$$

and the partition function obeys the recursion relation

$$Z_{L,N} = \sum_{k \geq 0} p_k Z_{L-1, N-k}. \quad (\text{A.14})$$

In order to make the link between the results of this section and those of Appendix A.1 one normalises the  $p_k$  as

$$\pi_k = \frac{p_k}{\sum_k p_k},$$

whenever the denominator is finite, thus recovering the probabilities  $\pi_k$  defined in (A.1). So doing, (A.12) is proportional to (A.5) and there is identity between (A.11)

and (A.6), (A.13) and (A.7) and (A.14) and (A.9). For instance the ‘balls-in-boxes’ model [1] has energy function

$$E(n_i) = \ln(n_i + 1),$$

yielding

$$p_k = \frac{1}{(1+k)^\beta}, \quad \pi_k = \frac{1}{\zeta(\beta)} \frac{1}{(1+k)^\beta},$$

where  $\sum_k p_k = \zeta(\beta)$  is the Riemann zeta-function. This model is the discrete counterpart of the case considered in the bulk of the paper where  $f_X(x)$  has a power-law tail (1.3). Here  $\pi_k \sim k^{-\beta}$ , with  $\beta$  playing the role of  $1 + \theta$ .

### Appendix A.3. Zero range process

*Definition.* The zero range process can be seen as a dynamical extension of the class of static urn models discussed above. We again consider a finite connected graph, made of  $L$  sites. At any time  $t$  a configuration of the system is specified by the values taken by the occupation numbers  $N_i(t)$ , now functions of time. The dynamics of the system consists in transferring a particle from the departure site with label  $d$ , containing  $N_d = k$  particles, to the arrival site with label  $a$  containing  $N_a = l$  particles. By definition of a ZRP, the transfer rate is

$$W(d, a, k) = w_{d,a} u_k,$$

where  $u_k$  only depends on the occupation  $N_d = k$  of the departure site and  $w_{d,a}$  accounts for diffusion from site  $d$  to site  $a$ . To simplify, let us restrict the discussion to diffusion processes such that the stationary state is uniform. The stationary probability of a configuration has the product form (A.11) where the factor  $p_k$  obeys the condition  $p_k u_k = p_{k-1}$ , which gives the explicit form

$$p_k = \frac{1}{u_1 \dots u_k}.$$

The statics of this ZRP is therefore the same as that of the urn model sharing the same  $p_k$ . Its partition function (A.12) obeys the recursion relation (A.14) and the stationary single-site occupation probability is given by (A.13).

Conversely, given an urn model, the corresponding ZRP has hopping rate  $u_k = p_{k-1}/p_k$ . For the balls-in-boxes model [1] this yields [3, 4]

$$u_k = \left(1 + \frac{1}{k}\right)^\beta \approx 1 + \frac{\beta}{k}.$$

*A prototypical condensing ZRP.* The model with hopping rate

$$u_k = 1 + \frac{b}{k} \tag{A.15}$$

is a well studied example of condensing ZRP. The weights  $p_k$  are given by

$$p_k = \frac{\Gamma(b+1) k!}{\Gamma(k+b+1)} = \int_0^1 du u^k b(1-u)^{b-1} \approx \frac{\Gamma(b+1)}{k^b},$$

with generating function

$$P(z) = \sum_{k \geq 0} z^k p_k = \int_0^1 du \frac{b(1-u)^{b-1}}{1-zu} = {}_2F_1(1, 1; b+1; z),$$

where  ${}_2F_1$  is the hypergeometric function. This function has a branch cut at  $z = z_c = 1$ , with a singular part of the form

$$P_{\text{sg}}(z) \approx A P(1)(1-z)^{b-1}, \quad A = \frac{(b-1)\pi}{\sin \pi b}$$

so that  $P(z)$  is only differentiable  $n \equiv \text{Int}(b) - 1$  many times at  $z = z_c = 1$ :

$$P(z) \approx P(1) + (z-1)P'(1) + \dots + \frac{(z-1)^n}{n!}P^{(n)}(1) + P_{\text{sg}}(z),$$

with

$$P(1) = \frac{b}{b-1}, \quad P'(1) = \frac{b}{(b-1)(b-2)}, \dots$$

In the thermodynamic limit ( $L \rightarrow \infty$  at fixed density  $N/L = \rho$ ), the system has a continuous phase transition at the critical density

$$\rho_c = \frac{P'(1)}{P(1)} = \frac{1}{b-2},$$

whenever  $b > 2$ . The critical density separates a fluid phase from a condensed phase. In the fluid phase ( $\rho < \rho_c$ ), the occupation probabilities  $f_k$  fall off exponentially. At the critical density ( $\rho = \rho_c$ ), they fall off as a power law:

$$f_k = \frac{p_k}{P(1)} \approx \frac{(b-1)\Gamma(b)}{k^b}. \quad (\text{A.16})$$

In the condensed phase ( $\rho > \rho_c$ ), for a large and finite system, the particles form a uniform critical background and a macroscopic condensate, consisting (on average) of  $\Delta$  excess particles with respect to the critical state, where

$$\Delta = N - L\rho_c = L(\rho - \rho_c).$$

The condensate appears as a hump in the stationary distribution  $f_k$ . The expression of the partition function  $Z_{L,N}$  deep in the condensed phase, i.e., for  $\Delta = L(\rho - \rho_c) \gg 1$  is [12]

$$Z_{L,N} \approx (b-1)\Gamma(b) \frac{L}{\Delta^b} P(1)^L. \quad (\text{A.17})$$

## Appendix B. Asymptotics of Hermite polynomials

We want to demonstrate (5.14) which holds if  $z > \sqrt{2n}$ . The generating function of the Hermite polynomials, defined as

$$H_n(z) = (-)^n e^{z^2} \frac{d^n}{dz^n} e^{-z^2},$$

is

$$\sum_{n \geq 0} H_n(z) \frac{u^n}{n!} = e^{2zu - u^2}.$$

Performing a saddle-point expansion of

$$H_n(z) = n! \int \frac{du}{2\pi i u} e^{2zu - u^2 - n \ln u},$$

one finds the saddle point

$$u_c = \frac{z - V}{2}, \quad V = \sqrt{z^2 - 2n},$$

finally yielding

$$H_n(z) \approx e^{(z^2 - zV - n)/2} (z + V)^n \sqrt{(1 + z/V)/2}, \quad (z > \sqrt{2n}),$$

which is (5.14).

## References

- [1] Bialas P, Burda Z and Johnston D 1997 Nucl. Phys. B **493** 505
- [2] Bialas P, Bogacz L, Burda Z and Johnston D 2000 Nucl. Phys. B **575** 599
- [3] Drouffe J M, Godrèche C and Camia F 1998 J. Phys. A **31** L19
- [4] Godrèche C 2007 Lect. Notes Phys. **716** 261
- [5] Spitzer F 1970 Advances in Math. **5** 246
- [6] Andjel E D 1982 Ann. Prob. **10** 525
- [7] Evans M R, 2000 Braz. J. Phys. **30** 42
- [8] Jeon I, March P and Pittel B 2000 Ann. Probab. **28** 1162
- [9] Godrèche C 2003 J. Phys. A **36** 6313
- [10] Grosskinsky S, Schütz G M and Spohn H 2003 J. Stat. Phys. **113** 389
- [11] Evans M R and Hanney T 2005 J. Phys. A **38** R195
- [12] Godrèche C and Luck J M 2005 J. Phys. A **38** 7215
- [13] Ferrari P A, Landim C and Sisko V V 2007 J. Stat. Phys. **128** 1153
- [14] Godrèche C and Luck J M 2012 J. Stat. Mech. P12013
- [15] Armendariz I and Loulakis M 2009 Probab. Theory Relat. Fields **145** 175
- [16] Armendariz I and Loulakis M 2011 Stochastic Process. Appl. **121** 1138
- [17] Armendariz I, Grosskinsky S and Loulakis M 2013 Stochastic Process. Appl. **123** 3466
- [18] Majumdar S N, Evans M R and Zia R K P 2005 Phys. Rev. Lett. **94** 180601
- [19] Evans M R, Majumdar S N and Zia R K P 2006 J. Stat. Phys. **123** 357
- [20] Evans M R and Majumdar S N 2008 J. Stat. Mech. P05004
- [21] Janson S 2012 Prob. Surveys **9** 103
- [22] Denisov D, Dieker A B and Shneer V 2008 Ann. Probab. **36** 1946
- [23] Chleboun P and Grosskinsky S 2010 J Stat Phys **140** 846
- [24] Zolotarev V M 1986 *One-dimensional stable distributions* Translations of Mathematical Monographs **65** (American Mathematical Society: Providence)
- [25] Gnedenko B and Kolmogorov A 1954 *Limit Distributions for Sums of Independent Random Variables* (Addison-Wesley: Cambridge, Mass)
- [26] Banderier C, Flajolet P, Schaeffer G and Soria M 2001 Random Struct. Alg. **19** 194 (John Wiley: New York)
- [27] Beltran J and Landim C 2012 Probab. Theor. Rel. Fields **152** 781
- [28] Landim C 2014 Comm. Math. Phys. **330** 1
- [29] Landim C 2018 arXiv:1807.04144
- [30] Armendariz I, Grosskinsky S and Loulakis M 2017 Probab. Theory Relat. Fields **169** 105.
- [31] Szavits-Nossan J, Evans M R and Majumdar S N 2014 PRL **112** 020602
- [32] Filiasi M, Livan G, Marsili M, Peressi M, Vesselli E and Zarinelli E 2004 J. Stat. Mech. P09030
- [33] Gradenigo G and Bertin E 2017 Entropy **19** 517
- [34] Corberi F 2015 J. Phys. A **48** 465003