Density of near-extreme events

Numerically generating density of states (DOS) and studying its limiting behaviour

Vishu Saini[1]

Manoj Kumar Singh[2]

Submitted to Prof. Dibyendu Das[3]

[1] *MSc Physics, II Semester (215120051), Department of Physics, IIT Bombay*
[2] *MSc Physics, II Semester (215120025), Department of Physics, IIT Bombay*
[3] *Professor, Department of Physics, IIT Bombay*

April 12, 2022

In this work, we study quantitatively the phenomenon of the crowding of events near the *extreme values* of *independent and identically distributed (IID)* random variables. We achieve this by computing the *density of states (DOS)* for these variables, which serve as a measure of "crowding of events" near extreme values. We will analyze the behavior of the density of states(DOS) by numerically computing its evolution with large values of N (number of IID variables). We will show how the mean DOS converges to three different limiting forms depending on whether the tail of the parent probability distribution of IID random variables decays slower than pure exponential, faster than pure exponential, or as a pure exponential function.

---

1 Introduction

Extreme value statistics (EVS) deals with the statistics of maximum or minimum values of a set of random observations. It is widely used in different fields like engineering, finance, studies of atmospheric phenomena, etc. In general we deal with the independent and identically distributed random variables and the statistics of these maximization and minimization are dealt using three well defined distributions 1.Frechet, 2.Gumbel and 3.Weibull. These three are used differently depending on the behaviour of the parent distribution. We use Frechet distribution when our probability distribution function p(X) has a tail which decays as a power law ($\sim x^{\alpha}$), Gumbel when the distribution decays faster than any power law ($> x^{\alpha}$) for any value of α and is unbounded and Weibull when tail is unbounded.
While the statistics of the maximas and minimas plays an important role in different fields, the behaviour of distribution near the maximas or minimas also play an important role in understanding whether the maxima that we are dealing with is alone at the top or there are also some nearby values lesser than the maxima. The study of these nearby values is quite useful as sometimes we can't measure the exact extrema of the distribution so the study of the behaviour of the distribution at the nearby points can sometimes provide a fairly good idea about the extremas. This concept is widely used in different problems like the study of level density for Bose-Einstein gas.

Here we will be dealing with only identical independent distributions as the study of even weakly correlated distributions will be quite difficult. Also the study of IID distributions give a fairly good idea even about weakly correlated distributions. So during this work we will be dealing only with IID variables.

2 Density of States

Consider a sequence with N IID random variables $\{X_1,X_2,....,X_N\}$. We will also consider here that they all are derived from a common probability distribution function $p(X)$. Let us take the maxima of the sequence to be X_{max}. We can write this as

$$X_{max} = max(X_1, X_2, ..., X_N) \tag{1}$$

Since we assumed X_1,X_2,...X_N to be IID variables but they do are correlated by the above relation.

Now let us consider a measure i.e, Density of states which describes the crowding of observation near maxima X_{max}.

$$\rho(r,N) = \frac{1}{N}\sum_{X_i \neq X_{max}}^{N-1} \delta[r - (X_{max} - X_i)] \tag{2}$$

where r describes how far X_{max} is from the point near which we are observing the crowding of events. X_i is the value of the observation for the i^{th} observation. Here we have not included the term of X_i=X_{max} because r is itself measured from X_{max}. Now consider

$$\int_0^\infty \rho(r,N)\,dr$$

$$= \int_0^\infty \frac{1}{N}\sum_{X_i\neq X_{max}}^{N-1} \delta[r-(X_{max}-X_i)]\,dr$$

$$= \sum_{X_i\neq X_{max}}^{N-1} \frac{1}{N}\int_0^\infty \delta[r-(X_{max}-X_i)]\,dr$$

$$= \frac{1}{N}\sum_{X_i\neq X_{max}}^{N-1} (1)$$

$$\therefore \int_0^\infty \rho(r,N)\,dr = 1-\frac{1}{N} \tag{3}$$

Since X_{max} >X_i for all i so for every X_{max} - X_i there must be some r between 0 to ∞ (as the separation between X_{max} and X_i is always finite) at which the argument of dirac function goes to 0.

Now let us take $\overline{\rho(r,N)}$ to be the mean DOS. It has a rich limiting behaviour as N $\rightarrow \infty$. So depending on the tail of the distribution $p(X)$, the governing distribution for $\overline{\rho(r,N)}$ is decided. So if the tail of the parent distribution $p(X)$ decreases slower than pure exponential function then the behaviour of $\overline{\rho(r,N)}$ is dealt using corresponding extreme value distribution. If the tail of $p(X)$ decreases faster than a pure exponential then it is related to the parent distribution itself (we will be proving this in this part) .[1] Atlast in case when $p(X)$ is a pure exponential then $\overline{\rho(r,N)}$ is dealt with something entirely different.

Now let us take X_{max} = x then rest of the (N-1) distributions are decided by the PDf $p(X)$. Let us further take common conditional probability function such that x still remains the maxima of the sequence

$$P_{cond}(X,x) = \frac{p(X)}{\int_{-\infty}^{x} p(y)\,dy} \tag{4}$$

where $p(X)$ is parent distribution function. We have taken the upper limit of the integral in the denominator as x because x is the maximum value of the sequence so no other X_i can exceed value x. Let us take $\overline{\rho_{cond}(r,N,x)}$ as the average density at some point r units away from x having N IID variables. So,

$$\overline{\rho_{cond}(r,N,x)}$$

$$= \int_0^\infty \frac{1}{N}\sum_{X_i\neq X_{max}}^{N-1} \delta[r-(X_{max}-X_i)]\,dr \cdot P_{cond}(x-r,x)$$

$$\therefore\ \overline{\rho_{cond}(r,N,x)} = \frac{N-1}{N}P_{cond}(x-r,x) \tag{5}$$

Using these we can write the PDF of the maximum value X_{max}=x

$$p_{max}(x,N) = Np(x)\left[\int_{-\infty}^{x} p(y)\,dy\right]^{N-1} \tag{6}$$

Here we have multiplied with N because any of the N IID variables can have maximum value as x. Now using the total probability theorem we can write

$$\overline{\rho(r,N)} = \int_{-\infty}^{\infty} \overline{\rho_{cond}(r,N,x)}\,p_{max}(x,N)\,dx \tag{7}$$

also then $\overline{\rho(r,N)}$

$$= \int_{-\infty}^{\infty} \frac{N-1}{N}p_{cond}(x-r,x)Np(X)\left[\int_{-\infty}^{x} p(y)\,dy\right]^{N-1} dx$$

$$= (N-1)\int_{-\infty}^{\infty} p_{cond}(x-r,x)p(x)\left[\int_{-\infty}^{x} p(y)\,dy\right]^{N-1} dx$$

$$= (N-1)\int_{-\infty}^{\infty} \frac{p(x-r)}{\int_{-\infty}^{x} p(x)\,dx}p(x)\left[\int_{-\infty}^{x} p(y)\,dy\right]^{N-1} dx$$

$$= (N-1)\int_{-\infty}^{\infty} p(x-r)p(x)\left[\int_{-\infty}^{x} p(y)\,dy\right]^{N-2} dx$$

$$\therefore \overline{\rho(r,N)} = \int_{-\infty}^{\infty} p(x-r)p_{max}(x,N-1)\,dx \tag{8}$$

This is an important result valid for all N.

2.1 Limiting behaviour of average DOS

Now we will try to analyse the limiting behaviour for large value of N. Since from extreme value statistics we know that for large N (i.e, $N \longrightarrow \infty$)

$$b_N P_{max}(x = a_N + b_N z) \longrightarrow f(z)$$

where $f(z)$ is universal and can belong to either of the three classes of the distributions Frechet, Gumbel or Weibull depending on the tail of $p(X)$. For example if $p(X) \sim exp(-X^\delta)$ for large X, then $a_N \sim(\ln N)^{1/\delta}$ and $b_N \sim\delta^{-1}(ln\mathrm{N})^{1/\delta-1}$. Then depending on different values of δ limiting value of b_N is decided and which further decides the functional form of $f(z)$. During next section we will be using equation (8)

2.1.1 Power-tail law

If $p(X)$ has slower than e^{-x} tail (i.e, $\delta < 1$), $b_N \rightarrow\infty$ as N $\rightarrow\infty$

$$b_N p(b_N + a_N - r) \longrightarrow \delta\left(z - \frac{r - a_N}{b_N}\right)$$

$$\implies \overline{\rho(r,N)} \longrightarrow \frac{1}{b_N} f\left(\frac{r - a_N}{b_N}\right) \tag{9}$$

2.1.2 Faster than power law

If $p(X)$ has faster than e^{-x} tail (i.e, $\delta > 1$), $b_N \rightarrow 0$ as N $\rightarrow\infty$

$$p_{max}(x,N) \longrightarrow \delta(x - a_N)$$

$$\implies \overline{\rho(r,N)} \longrightarrow p(a_N - r) \tag{10}$$

2.2 Average DOS depending on three Extreme Value Distributions

Now let us consider the three cases of Frechet, Gumbel and Weibull one by one depending on the tail of the parent distribution $p(X)$.

2.2.1 power-law tail

In case when the tail decreases as a power law then p(X) is given by

$$p(X) = \frac{\alpha exp(-X^{-\alpha})}{X^{1+\alpha}}, \quad \alpha > 0, \quad X \in [0, \infty)$$

then in this case $a_N = 0$ and $b_N \rightarrow N^{\frac{1}{\alpha}} \rightarrow\infty$ as N $\rightarrow\infty$.

$$\implies \overline{\rho(r,N)} \longrightarrow \frac{1}{b_N} f\left(\frac{r}{b_N}\right) \tag{11}$$

where $f(z)$ will be decided by Frechet distribution,

$$f(z) \longrightarrow f_1(z) = \frac{\alpha exp[-z^{-\alpha}]}{z^{1+\alpha}}, z > 0.$$

2.2.2 Faster than power-law, but unbounded tail

In case when tail varies faster than any power law then,

$$p(X) = \delta X^{\delta-1} exp(-X^{\delta})$$

then $a_N = (lnN)^{\frac{1}{\delta}}$ and $b_N \rightarrow \delta^{-1}(lnN)^{\frac{1}{\delta-1}}$ when N $\rightarrow \infty$. So depending on the value of δ, value of b_N will be decided.

• If $\delta < 1$ then $b_N \rightarrow\infty$ then

$$\implies \overline{\rho(r,N)} \longrightarrow \frac{1}{b_N} f\left(\frac{r - a_N}{b_N}\right) \tag{12}$$

where f is given by Gumbel distribution,

$$f_2(z) = e^{(-z - e^{(-z)})}$$

• If $\delta = 1$ then $b_N \rightarrow 1$ then

$$\implies \overline{\rho(r,N)} \longrightarrow g(r - a_N) \tag{13}$$

here in this borderline case ($\delta = 1$), g is given by completely different scaling function,

$$g(z) = e^z[1 - (1 + e^{-z})e^{-e^{-z}}]$$

• If $\delta > 1$ then,

$$\implies \overline{\rho(r,N)} \longrightarrow p(a_N - r) \tag{14}$$

where p(X) is given by the parent distribution.

2.2.3 Bounded tail

In case of a bounded tail,

$$p(X) = \beta a^{-\beta}(a - X)^{\beta-1}, \quad X \in [0, a]$$

then a_N = a and $b_N \rightarrow$ a $N^{\frac{-1}{\beta}} \longrightarrow 0$ as $N \rightarrow \infty$

$$\implies \overline{\rho(r,N)} \longrightarrow p(a_N - r) \tag{15}$$

where p(X) is the parent distribution.

Hence we have explicitly shown one by one how the average DOS will be decided depending on the tail of the parent distribution function $p(X)$.

3 Numerical generation of DOS

We will now show how $\overline{\rho(r,N)}$ evolves as N takes larger and larger values by taking different forms of parent probability distribution function $p(X)$.

3.1 Analysis Methodology

We use the following steps to numerically compute $\overline{\rho(r,N)}$:

- First define explicit form of $p(X)$, using required parameters. (Here we work with three forms as discussed in Section 2.2)
- Determine scale factors a_N and b_N. (again as determined in Section 2.2)
- Determine $p_{max}(x,N)$ by numerically integrating equation (6); i.e

$$p_{max}(x,N) = Np(x)[\int_{-\infty}^{x} p(y)\, dy]^{N-1} \tag{16}$$

- Determine $\overline{\rho(r,N)}$ by numerically integrating equation (8); i.e

$$\overline{\rho(r,N)} = \int_{-\infty}^{\infty} p(x-r)p_{max}(x,N-1)\,dx \quad (17)$$

- Plot $\overline{\rho(r,N)}$ for various values of N and $f(z)$ using appropriate scale transformations involving a_N and b_N.

4 Results

4.1 Power-law tail

Explicit form of $p(X)$ used is:

$$p(X) = \frac{\alpha\exp(-X^{-\alpha})}{X^{1+\alpha}}, \quad \alpha > 0, \quad X \in [0,\infty) \quad (18)$$

with $a_N = 0$ and $b_N = N^{1/\alpha}$

Following the steps as discussed in section of Analysis methodology; we have computed $\overline{\rho(r,N)}$ for two different values of parameter α, and traced its evolution with increasing values of N. $b_N\overline{\rho(r,N)}$ vs r/b_N curves are plotted in Figures 1 and 2. "limit" implies $N \longrightarrow \infty$ limit of $b_N\overline{\rho(r,N)}$ i.e $f(r/b_N)$, where $f(z)$ belongs to *Frechet class*.

We observe that $b_N\overline{\rho(r,N)}$ converges to $f(r/b_N)$, much faster for $\alpha = 2$ then for $\alpha = 5$.

4.2 Faster than power law, but unbounded tail

Form of $p(X)$ in this case is given by

$$p(X) = \delta X^{\delta-1}\exp(-X^{\delta}), \quad \delta \leq 1, \quad X \in [0,\infty) \quad (19)$$

With $a_N = (\ln N)^{1/\delta}$ and $b_N = \delta^{-1}(\ln N)^{1/\delta-1}$.

We have computed two different cases of δ.

Case I: $\delta < 1$ In this case $b_N\overline{\rho(r,N)}$ converges to $f(\frac{r-a_N}{b_N})$ as shown in Figure 3 and Figure 4, where $f(z$ belongs to *Gumbel class*. Here, "limit" represents $f(\frac{r-a_N}{b_N})$. For both $\delta = 1/2$ and $\delta = 1/3$ convergence to limit is of similar order.

Case II: $\delta = 1$ In this case, $\overline{\rho(r,N)}$ converges to $g(r-a_N)$ as shown in Figure 5. Here, "limit" represents $g(r-a_N)$. The convergence is much faster than in $\delta < 1$ case.

4.3 Bounded tail

In this case, $p(X)$ has the form;

$$p(X) = \beta a^{\beta}(a-X)^{\beta-1}, \quad \beta > 0, \quad X \in [0,a] \quad (20)$$

With $a_N = a$ and $b_N = aN^{-1/\beta}$.

We have computed $\overline{\rho(r,N)}$ for $a = 10$ and $\beta = 3/2$ in Figure 6. In this case, $\overline{\rho(r,N)}$ converges to $p(a_N - r)$. We observe that $\overline{\rho(r,N)}$ starts converging to $p(a_N - r)$ as values of N starts increasing.

5 Conclusion

We computed mean DOS by numerically integrating equations (16) and (17) and studied its limiting behaviour for large values of N. Three different forms of parent distribution function $p(X)$ ((18), (19) and (20)) were taken and mean DOS was computed and plotted for each case. In case of *Power-law tail* (18), we see that mean DOS converges to *Frechet class* for large N. However, mean DOS converges faster for $\alpha = 2$, than for for $\alpha = 5$. In case of *Faster than Power-law but unbounded tail* (19), we see mean DOS converges to *Gumbel class* for large N. It converges approximately in a similar fashion for $\delta = 1/2$ and $\delta = 1/3$, but convergence is slow as compared to previous case. However, for $\delta = 1$, the pure exponential case, it converges to $g(r-a_N)$ extremely faster than any other case. In case of *Bounded tail* (20) , we see that mean DOS converges to $p(a_n - r)$.

6 Acknowledgments

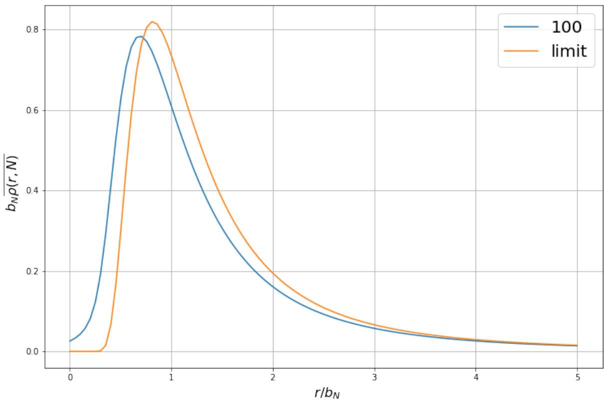
We express our deep gratitude to our course instructor and advisor for this project, Prof. Dibyendu Das, for allowing us to work on this project under his guidance. We are highly indebted to him for his commitment and dedication to the course as well as to the students. It would not have been possible for us to complete this work without his key insights and supervision that guided us and provided us with the much-needed direction. It has been a great privilege to be a student of his; we owe him for the immense ways he has shaped us thus far.

References

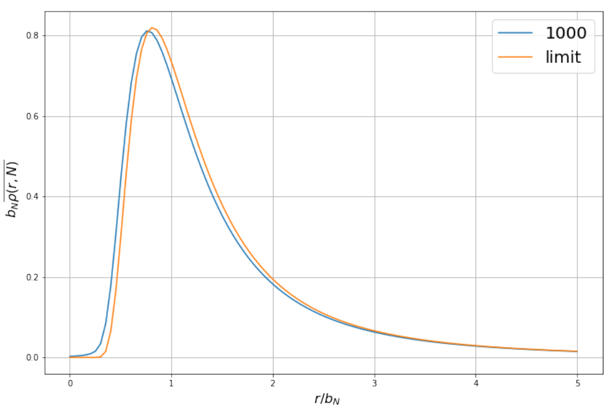
[1] Sanjib Sabhapandit, Satya N. Majumdar. Phys. Rev. Lett. vol. 98, pp. 140201 (2007). arXiv:cond-mat/0701375 [cond-mat.stat-mech]

[2] Front. Phys., 10 December 2020. REVIEW article. *The Three Extreme Value Distributions: An Introductory Review* https://doi.org/10.3389/fphy.2020.604053

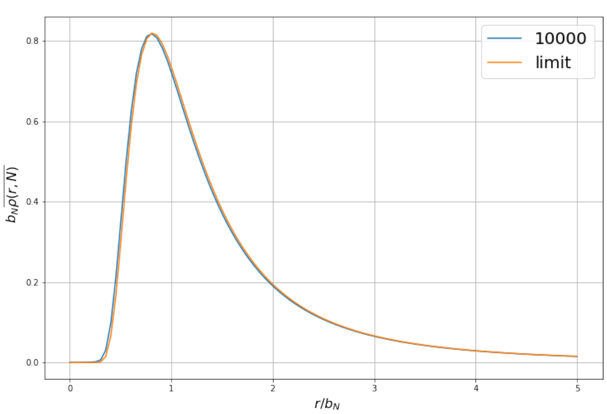
[3] Sanjib Sabhapandit and Satya N. Majumdar Laboratoire de Physique The´orique et Mode‘les Statistiques (UMR 8626 du CNRS), Universite´ Paris-Sud, Ba^timent 100, 91405 Orsay Cedex, France. *Density of Near-Extreme Events*



(a)

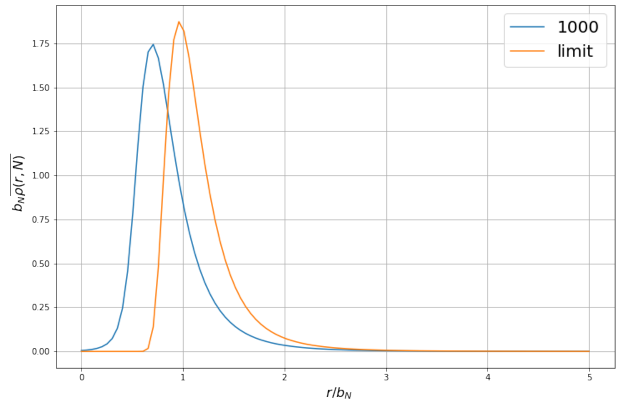


(b)

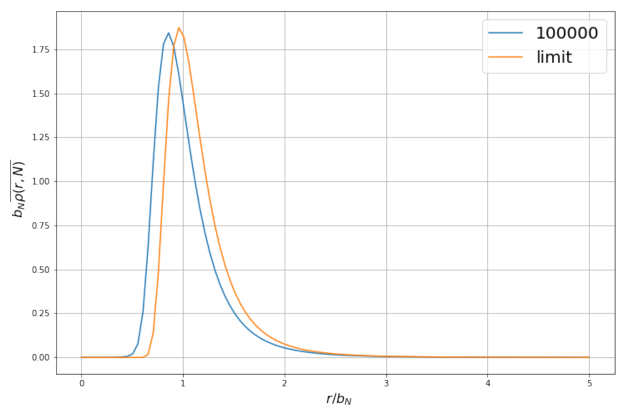


(c)

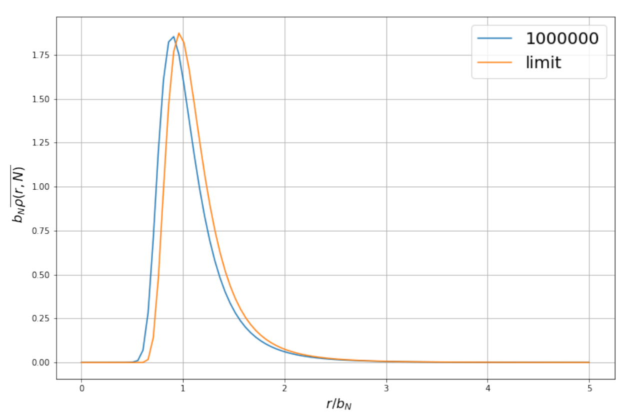
Figure 1: Power law tail : for $\alpha = 2$



(a)

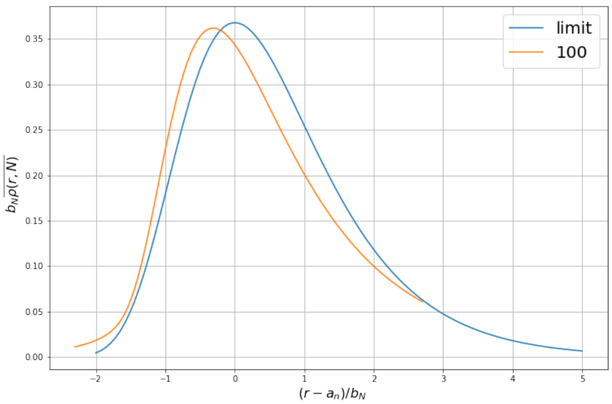


(b)

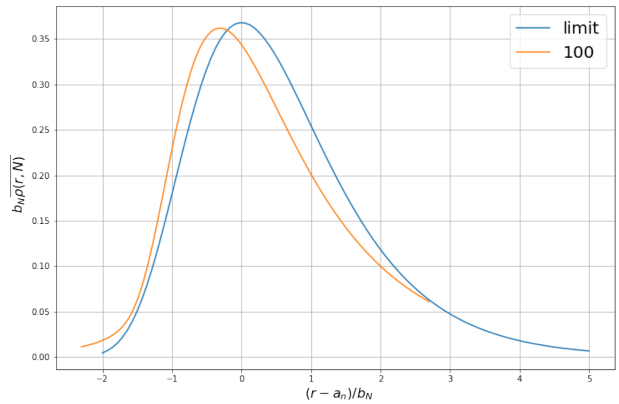


(c)

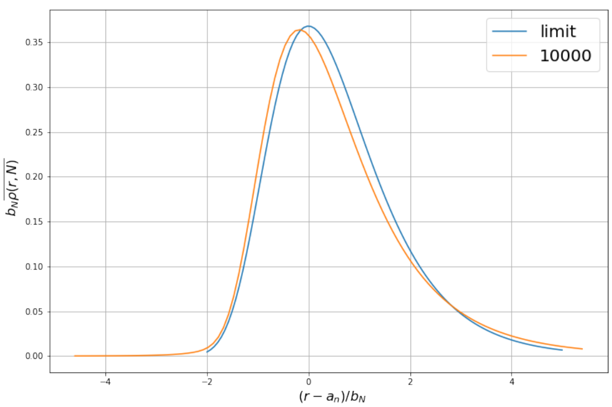
Figure 2: Power law tail : for $\alpha = 5$



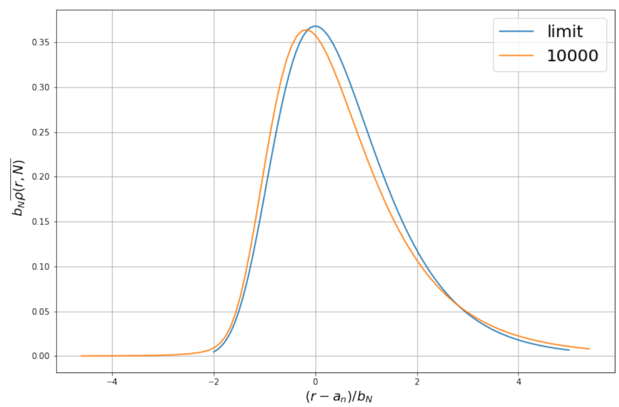
(a)



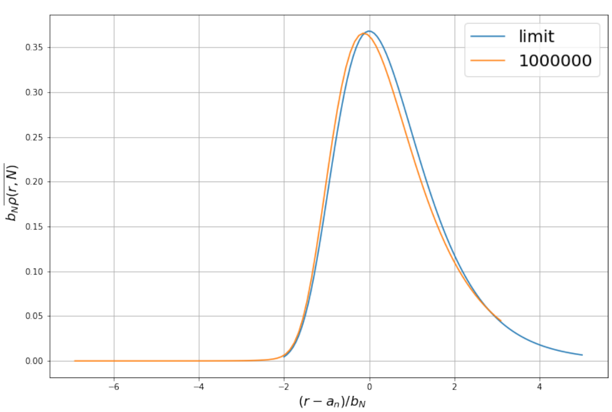
(a)



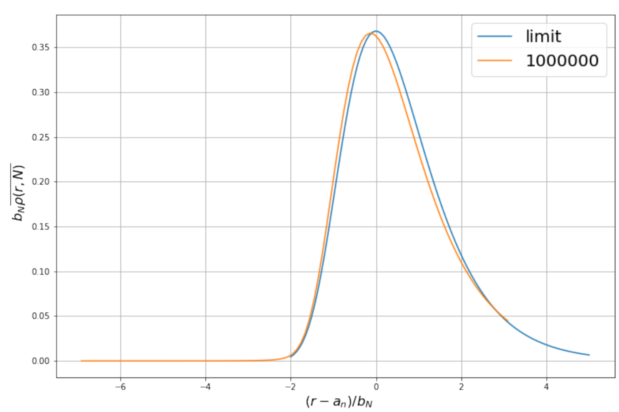
(b)



(b)



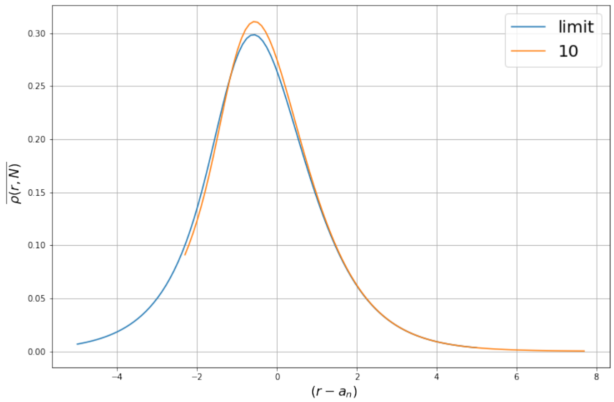
(c)



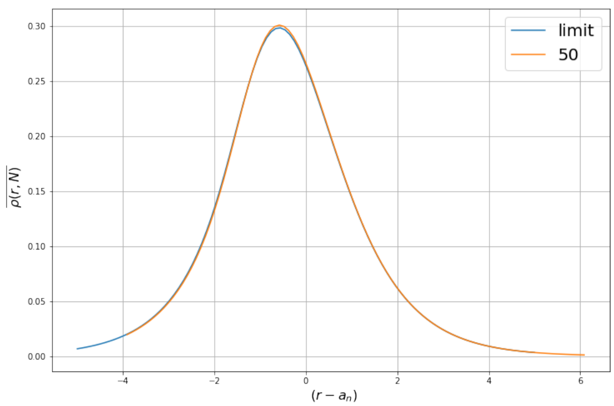
(c)

Figure 3: Faster than power law but unbounded tail : for $\delta = 1/2$

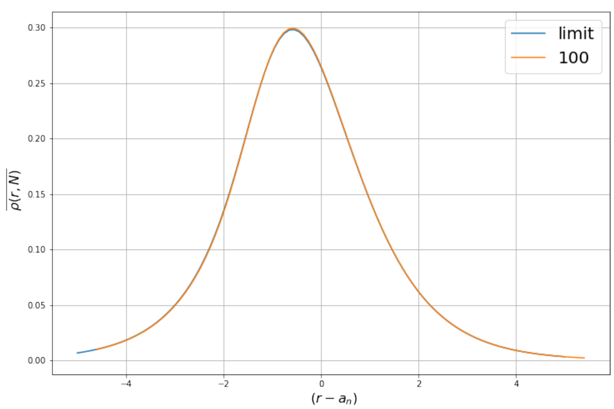
Figure 4: Faster than power law but unbounded tail : for $\delta = 1/3$



(a)



(b)



(c)

Figure 5: Pure exponential tail: for $\delta = 1$

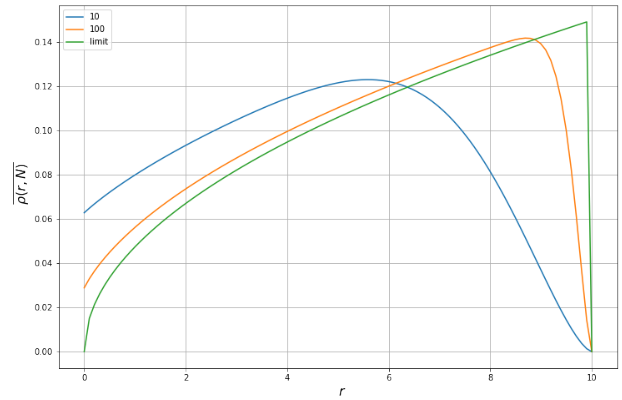


Figure 6: Slower than pure exponential tail: for a = 10

Code for Numerical Analysis

The code for numerical generation of mean DOS was written in Python, (Jupyter-notebooks) and is available on google drive link: Drive link