Analysis of the convergence of Monte Carlo averages

Alejandro Llorente Pinto
Supervised by Alberto Suárez González

Master of Science Thesis

Escuela Politécnica Superior
Universidad Autónoma de Madrid

April 2, 2012

Abstract

The weak and strong laws of large numbers and the central limit theorem describe the asymptotic convergence of sums of random variables in the limit of large samples. For small samples, empirical estimates of expected values can exhibit a behavior that strongly deviates from what is predicted by these theorems. In particular, the mean and the mode of the sample average could be very far apart. As a consequence, typical values of this average could be very different from its expected value. To analyze the convergence of the typical to the expected behavior, we propose to use the bias of the logarithm of the estimator. This bias is non-negative and decreases monotonically with the size of the sample.

In the literature, one finds different kinds of analysis of the convergence of sample averages to expected values as a function of sample size. Kagan and Nagaev (2001) propose a criterion to determine the largest moment order that can be consistently estimated from a sample of size M. The requirement of consistency is very stringent and leads to extremely low bounds. As an alternative, we extend the analysis of the random energy model (Derrida,1981) to the convergence of sample averages. The REM is a simplified model of disordered systems, whose energy levels are independent random variables sampled from a Gaussian distribution. The quantity of interest is the partition function. In the standard variant of REM, the partition function is the sum of M lognormal random variables whose shape parameter is proportional to the temperature and to $\log M$. From the partition function, one can define a *quenched* free energy. This quantity exhibits a second order phase transition as a function of temperature. Since the derivations of the REM are not entirely formal, we also review Large Deviation Theory (LDT), a branch of statistics that studies the probability of rare events. This theory can be used to provide rigorous derivations of the results of the REM analysis.

The main contribution of the current investigation is the application of the REM and LDT analysis to the problem of estimating the moment of order q from a sample of size M. In particular, we have analyzed the convergence of the sample moments of the lognormal, log-exponential-power-law, folded normal, Weibull and Chi-squared distributions. In all these cases, the moment estimator exhibits a phase transition in the limit of large samples ($M \to \infty$) and large moment orders ($q \to \infty$) with $\log M/q \to constant$. This means that for a given moment order, q, there is a a threshold $M_c(q)$ that marks an abrupt transition in the behavior of the empirical moment estimator. For small samples ($M < M_c(q)$), the logarithm of the estimator is biased. For samples whose size is larger than $M_c(q)$, the estimator is asymptotically unbiased. We also analyze the Pareto distribution, which is out of the scope of REM-LDT. In this case, the logarithm of the empirical moment estimator does not exhibit a phase transition and is always biased. Yet, the dependence of this bias with the size of the sample can be accurately approximated using right-censored moments with an appropriate choice of the truncation threshold (e.g. the mode or the mean of the sample maximum).

Finally, the validity of the analysis performed is illustrated with the results of extensive Monte Carlo simulations.

Contents

Chapter 1

Introduction

Monte Carlo (MC) methods are a family of numerical techniques that make use of random numbers to obtain the solution of problems of different nature. Even though the generation of random numbers is the defining characteristic of these methods, the problems addressed need not be stochastic. An attractive feature of MC algorithms is that they are general methods that can provide accurate approximations when analytical or deterministic numerical approaches fail or are difficult to implement. Furthermore, they take can be adapted to advantage of the large processing capacity of current computational systems to address problems in a wide range of areas of application. For instance, in the health field, MC has been applied to model stochastic processes in nuclear medicine [1]; in engineering, MC techniques have been used for the analysis of industrial processes [2]; in finance, they have been used to value financial options and derivatives [3, 4]; they have been also applied to problems in protein folding [5]; in chemistry, to model molecular dynamics [6]; and in physics they have been used for studying phase transitions [7].

A common step in most of applications of MC methods is the estimation of expected values of functions of random variables. Consider the random variable X, characterized by the probability density function $f_X(x)$. The expected value of $g(x)$ with respect to this distribution is

$$\mathbb{E}[g(X)] = \int g(x) f_X(x) dx. \tag{1.1}$$

Given $X_1, ..., X_M$ a random sample distributed as X, the sample average

$$Z_M[g] = \frac{1}{M} \sum_{i=1}^{M} g(X_i). \tag{1.2}$$

is an empirical estimator of the expected value $\mathbb{E}[g(X)]$. This estimator is unbiased for any value of M

$$\mathbb{E}\left[Z_M[g]\right] = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[g(X_i)] = \mathbb{E}\left[g(X)\right] \tag{1.3}$$

because of the linearity of the expected value operator. Note that it is not necessary to assume independence to derive this property. Therefore, estimates obtained using Markov Chain Monte Carlo (MCMC) are also unbiased [8].

MC methods can also be used to approximate numerical quadratures. Consider the quadrature

$$I = \int_a^b g(x)dx. \tag{1.4}$$

The key of the application of MC techniques to the numerical estimation of I is to transform this quadrature into an expected value over a particular probability density function as in (1.1). In particular, multiplying and dividing by the length of the integration interval, one obtains

$$I = \int_a^b g(x)dx = (b-a)\int_a^b \frac{1}{b-a} g(x)dx. \tag{1.5}$$

Making the observation that the factor $f_X(x) = 1/(b-a)$ inside the quadrature can be interpreted as the density of $U[a,b]$, the uniform distribution in $[a,b]$, the quadrature (1.4) is approximated by

$$\widehat{I} = \frac{b-a}{M}\sum_{i=1}^{M} g(x_i) \tag{1.6}$$

where $\{x_i\}$ is a random sample of M values that are uniformly distributed in $[a,b]$. Taking the expected value of this expression, one obtains

$$\mathbb{E}[\widehat{I}] = \frac{b-a}{M}\sum_{i=1}^{M}\mathbb{E}\left[g(x_i)\right] \underset{i.d.}{=} (b-a)\mathbb{E}\left[g(x)\right] =$$
$$= (b-a)\int_a^b g(x)\frac{1}{b-a}dx = \int_a^b g(x)dx,$$

which is the quantity that we are after.

Moment estimation, which is one of the objects of our analysis, is a particular case of (1.1) when $g(x) = x^q$. Let $Z_M^{(q)}$ be the MC estimate of the qth moment from a sample of size M

$$Z_M[g] \overset{g(x)=x^q}{\equiv} Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q \tag{1.7}$$

Despite the fact that estimator is unbiased, as demonstrated in (1.3), nothing is said about how the MC average converges to the actual mean. The asymptotic convergence of the estimator to its expected value is described by the laws of Large Numbers:

Theorem 1.0.1 *(Weak Law of Large Numbers, WLLN) Let $\{X_i\}_{i=1}^M$ be a sequence of independent identically distributed (i.i.d.) random variables as X. We define $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$, then*

$$\lim_{M\to\infty}\mathbb{P}\left(|Z_M - \mathbb{E}[X]| > \delta\right) = 0 \underset{def}{\Longleftrightarrow} Z_M \xrightarrow{P} \mathbb{E}[X]$$

for any $\delta > 0$

The weak Law states that the probability of the event $|Z_M - \mathbb{E}[X]| > \delta$ decreases to zero as the sample size increases. However, for a sample of finite size, this property need not hold with a non-vanishing probability. Since it is desirable to ensure that the error is bound for samples of size greater than a finite value M_0, a more restrictive type convergence is needed. The *strong* law of large numbers is a translation of this more stringent requirement.

Theorem 1.0.2 *(The strong law of large numbers, SLLN) Let $\{X_i\}_{i=1}^M$ be a sequence of independent identically distributed (i.i.d.) random variables distributed as $f_X(x)$. Define $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$. Using these definitions, the strong law of large numbers states*

$$\mathbb{P}\left(\lim_{M\to\infty} Z_M = \mathbb{E}[X]\right) = 1 \underset{def}{\iff} Z_M \xrightarrow{a.s.} \mathbb{E}[X]$$

This type of convergence means that $|Z_M - \mathbb{E}[X]| > \delta$ only happens a finite number of times. That is, there exists an M_0 such that $|Z_M - \mathbb{E}[X]| < \delta, \forall M > M_0$ with probability 1.

Both the weak and the strong laws of large numbers are statements about the limit of sums of random variables. However, they do not provide any information on the distribution of Z_M in the limit $M \to \infty$. For this, one needs the central limit theorem, which describes the asymptotic form of the distribution of Z_M for random variables whose variance is finite.

Theorem 1.0.3 *(The central limit theorem (CLT)) Let $\{X_1, ..., X_M\}$ be a sequence of M i.i.d. random variables with a finite expected value $\mu < \infty$ and a finite variance $\sigma^2 < \infty$.*
Define $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$. The central limit theorem states

$$Z_M \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{M}\right) \tag{1.8}$$

where convergence in distribution is equivalent to convergence of the cdf of the random variables, that is,

$$\lim_{M\to\infty} F_{Z_M}(x) = \Phi(x; \mu, \sigma^2) \tag{1.9}$$

in the points where $F_{Z_M}(x)$ is continuous and $\Phi(x; \mu, \sigma^2)$ is the cdf of the Gaussian distribution with mean μ and variance σ^2.

Some observations can be made from this statement:

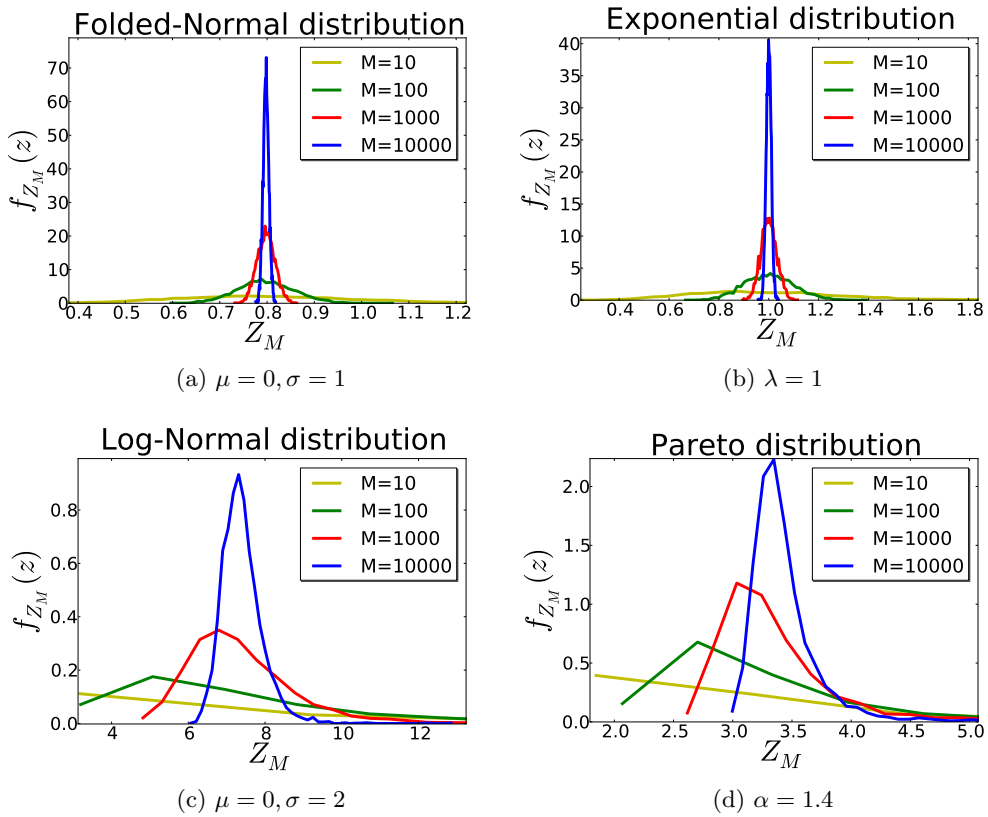
(i) the average of a sample of i.i.d. random variables tends to the distribution mean,

(ii) the variance tends to zero as the number of samples increases and

(iii) the convergence to a normal is independent of the distribution, provided that the variance is finite.

Since the CLT describes the asymptotic distribution of the sum of random variables, it is reasonable to ask the question of how many samples are necessary for the normal distribution that appears in this theorem to be an accurate approximation of the distribution of the MC average.

In spite of the general asymptotic validity of the CLT for sums of independent random variables of finite variance, the rate of convergence to the normal asymptotic behavior is distribution dependent. In his study on the convergence of MC averages of power-law distributions [9], Crovella identifies convergence patterns in averages of heavy-tailed data that are qualitatively different from the asymptotic regime described by the CLT, even when both the mean of the variance of the individual samples are finite (figure 1.1). For heavy-tailed distributions, such as the Pareto and the lognormal distributions, and for samples of small sizes, the mode of the empirical average is shifted with respect to the limiting value. By contrast, for distributions with exponentially decaying tails, such as the Gaussian or the exponential distribution, the mode of the MC estimate is close to the actual average even for relatively small samples. All these observations lead us to believe that there are *regimes* in which the finite size effects, such as the asymmetry of the distribution or shift of the mode with respect to the mean, cannot be accounted for by the CLT.



(a) $\mu = 0, \sigma = 1$

(b) $\lambda = 1$

(c) $\mu = 0, \sigma = 2$

(d) $\alpha = 1.4$

Figure 1.1: Distribution of the mean estimator as a function of sample size for different types of distributions.

The observed behavior does not depend on the bias of the estimate because, as shown

earlier, this bias is zero regardless of the distribution. From these observations, it is clear the CLT is not an accurate description for small samples. Specifically, the distribution of the sum of random variables is asymmetric. Furthermore, unlike for the Gaussian distribution, the mode does not coincide with the mean. Note that this discrepancy does not depend on the shape or the support of the probability distribution of the random variables in the sum. It seems to be a consequence solely of the tail properties of the probability distribution.

One might expect that there is a smooth transition between a small sample regime, in which finite size effects such as those observed in sums of heavy-tailed random variables, are dominant and, an asymptotic (large sample) regime in which the CLT provides an accurate approximation to the distribution of the empirical averages. In this work we show that for sufficiently heavy-tailed data, the transition is rather abrupt and can be approximately described in a certain limit as a second order phase transition in the bias of the *logarithm* of the estimator. In this limit this bias exhibits a discontinuous transition at a critical sample size M_c from a region of non-zero bias (small samples, $M < M_c$), to a regime in which it is unbiased (large samples, $M > M_c$) that is similar to the second order phase transition in the *quenched* free energy in the Random Energy Model [10]. The transition does not take place when the decay is algebraic: the bias of the logarithm of the estimator is different from zero for any sample size. The existence of an abrupt change in the behavior of the logarithm of the estimator as a function of sample is quite different from the smooth asymptotic decrease of the variance predicted by the CLT.

1.1 Mathematical formulation

In this section, we set up a framework to analyze the convergence of the MC estimator of the mean of non-negative random variables as a function of the size of the sample. We propose to use the bias of the logarithm of the estimator as a measure of the quality of the estimate. In the limit of large samples this bias is a measure of the discrepancy between the mode and the mean of the sample estimate. Since the mode represents the most probable value of the empirical estimator, this quantity quantifies how different are the typical and the expected values of the sample average. The expected value of this bias is shown to be monotonically decreasing with the sample size. Furthermore, for non-algebraically decaying heavy-tailed distributions, this expected value exhibits a second order phase transition similar to the one exhibited by the quenched free energy in the random energy model, a simplified model for glasses [10].

Consider a random variable X whose distribution function (cdf) is $F_X(x)$. The corresponding probability density function (pdf) is $f_X(x)$. The expected value of X is

$$\mathbb{E}[X] = \int_A x f_X(x) dx \tag{1.10}$$

where $A = \{x \in \mathbb{R} \ : f_X(x) \neq 0\}$. We focus on positive random variables, $A \subseteq \mathbb{R}^+$. By (1.3), $Z_M = \frac{1}{M} \sum_{i=1}^{M} X_i$ is an unbiased estimator of $\mathbb{E}[X]$. However non-linear functions of Z_M will in general be biased. In particular, consider bias of the logarithm of Z_M

$$B_M = \log \mathbb{E}[X] - \log Z_M \tag{1.11}$$

B_M is a random variable whose distribution depends on the sample size M. It is finite because the random variables $\{X_i\}$ are non-negative and the event $X_i = 0$ has measure zero.

The expected value of B_M is a measure of how different the typical and the average values of Z_M are, for large M. To understand why this is the case, let $f_M(z)$ the density function of Z_M. As M increases $f_M(z)$ becomes more concentrated around its mode z_M^*. Since $\log z$ varies only smoothly in the region where $f_M(z)$ is peaked , in the limit $M \to \infty$, the expectation of $\log Z_M$ is approximately

$$\mathbb{E}\left[\log Z_M\right] = \int dz f_M(z) \log z \approx \log z_M^* \tag{1.12}$$

Therefore,

$$\mathbb{E}[B_M] = \log \mathbb{E}[X] - \mathbb{E}[\log Z_M] \approx \log \mathbb{E}[X] - \log z_M^*, \tag{1.13}$$

which is a measure of the difference between the mode and the mean of $f_M(z)$.

We now state some properties of B_M. The first one is that it converges to zero as the sample size increases. The second one is that the convergence of its average to zero is monotonic.

Proposition 1.1.1 *The estimator* $\log Z_M$ *converges almost surely, in probability and in distribution to* $\log \mathbb{E}[X]$.

Proof: This property is a direct application of the continuous mapping theorem [11].

Theorem 1.1.2 *Let* $\{Z_M\}$ *defined on a metric space. If* h *is a function from this metric space* S *and the set of points of discontinuity has zero measure then*

1. $Z_M \xrightarrow{d} Z \Rightarrow h(Z_M) \xrightarrow{d} h(Z)$
2. $Z_M \xrightarrow{P} Z \Rightarrow h(Z_M) \xrightarrow{P} h(Z)$
3. $Z_M \xrightarrow{as} Z \Rightarrow h(Z_M) \xrightarrow{as} h(Z)$

By application of the Strong Law of Large Numbers we have almost sure convergence of Z_M to $\mathbb{E}[X]$. Then, using the continuous mapping theorem, we have almost-sure convergence of $\log Z_M$ to $\log \mathbb{E}[X]$. Note that almost-sure convergence implies convergence in probability. Finally, convergence in probability implies convergence in distribution, which completes the proof. □

Theorem 1.1.3 *The expectation of the bias of the logarithm is a monotonically decreasing function of* M

$$\mathbb{E}[B_M] \geq \mathbb{E}[B_{M+1}]$$

Proof: The derivation makes use of a theorem whose proof is given in appendix F. Before formulating the theorem, we define the quenched average as

$$\langle Y \rangle_g^M = \mathbb{E}\left[g^{-1}\left(\frac{1}{M} \sum_{i=1}^{M} g(Y_i) \right) \right] \tag{1.14}$$

where g is a convex function.

Theorem 1.1.4 *For a sequence of iid random variables* $\{Y_1, Y_2, \ldots, Y_M, Y_{M+1}, \ldots\}$

$$\langle Y \rangle_{M+1}^{g} \geq \langle Y \rangle_{M}^{g} \tag{1.15}$$

In our investigation we choose $g(x) = e^x$. Making the change of variables $X_i = e^{Y_i}$ in $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$ (which can be done because $X_i > 0,\ a.s.$) and applying theorem 1.1.4,

$$\begin{gathered}
\langle Y \rangle_{M+1}^{g} - \langle Y \rangle_{M}^{g} \geq 0 \\
\mathbb{E}[\log Z_{M+1}] - \mathbb{E}[\log Z_M] = \mathbb{E}[\log Z_{M+1}] - \mathbb{E}[\log Z_M] \pm \log(\mathbb{E}[X]) = \\
= (\log(\mathbb{E}[X]) - \mathbb{E}[\log Z_M]) - (\log(\mathbb{E}[X]) - \mathbb{E}[\log Z_{M+1}]) = \\
= \mathbb{E}[B_M] - \mathbb{E}[B_{M+1}] \geq 0
\end{gathered}$$

□

Finally one can also derive the following result for the convergence of the expectation of the estimator

Proposition 1.1.5

$$\lim_{M\to\infty} \mathbb{E}[\log Z_M] = \log \mathbb{E}[X]$$

Proof: Note that, by virtue of the Strong Law of Large Numbers, we have

$$\lim_{M\to\infty} Z_M = \mathbb{E}[X] \quad a.s. \tag{1.16}$$

If it were possible interchange the expected value and the limit in the left-side of the proposition the result would be true.
As we have seen in the previous property $\mathbb{E}[B_M] \geq \mathbb{E}[B_{M+1}]$. Therefore,

$$\mathbb{E}[\log Z_{M+1}] \geq \mathbb{E}[\log Z_M]$$

Finally

$$\lim_{M\to\infty} \mathbb{E}[\log Z_M] = \mathbb{E}\left[\lim_{M\to\infty} \log Z_M\right] = E[\log \mathbb{E}[X]] = \log \mathbb{E}[X]$$

□

Since $\mathbb{E}[B_M]$ is a monotonic decreasing function in M, $\mathbb{E}[B_1] \geq \mathbb{E}[B_M]\ \ \forall M \geq 1$. Therefore the maximum expected bias of the logarithm of the estimator is

$$B_{max} = \mathbb{E}[B_1] = \log \mathbb{E}[X] - \mathbb{E}[\log X] \tag{1.17}$$

Finally, we can define the normalized bias of the logarithm of Z_M

$$\widehat{B}_M = B_M / B_{max} \tag{1.18}$$

where $0 \leq \mathbb{E}[\widehat{B}_M] \leq 1$.

The study of the convergence of sample averages can be useful to identify different regimes. Namely, a regime in which neither the sample mean nor the sample variance have converged; a regime in which the mean has converged but the variance has not and, finally a regime in which both quantities have converged. In fact, the analysis in terms of the bias B_M can be extended to an arbitrary moment of the distribution by considering the mean of a random variable $Y = X^q$,

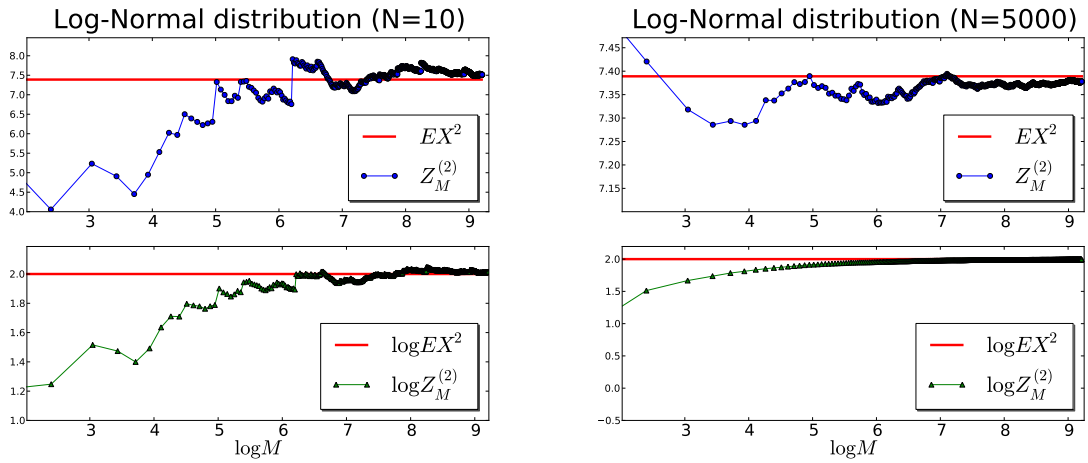
$$F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}(X < y^{1/q}) = F_X(y^{1/q}) \tag{1.19}$$

$$f_Y(y) = \frac{1}{q} f_X(y^{1/q}) y^{\frac{1-q}{q}} \tag{1.20}$$

Figure 1.2 illustrates the convergence of the simulated $Z_M^{(q)}$ and $\log Z_M^{(q)}$ with $\mathbb{E}[X^q]$ and $\log \mathbb{E}[X^q]$ respectively, for $q = 2, 4$, where X is a lognormal random variable with $\mu = 0$ and $\sigma = 1$. The monotonicity of the expected value of the bias logarithm of the sample average, proved in 1.1.4, is observed in this figure for both moments. By contrast, the bias of Z_M has a very irregular behavior with large jumps followed by a smoother decay.

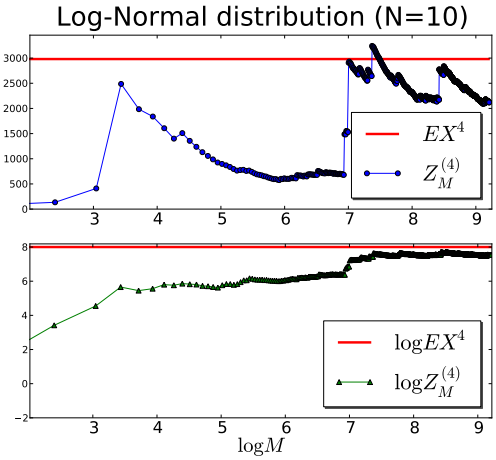
On the other hand, while in figure 1.2b the difference between $\log Z_M^{(2)}$ and $\log \mathbb{E}[X^2]$ becomes vanishingly small around a critical point $\log M \approx 7$. For the fourth order moment (figure 1.2d) this critical point is not reached yet. This illustrates the different properties of the empirical estimates of the moments of a random variable and the dependence of the rate of convergence on the order of the moment. This dependence is reasonable because increasing the the order of the moment makes the extreme values more dominant in the sample average sum. Thus, larger sample sizes are required for convergence. Our main goal is to derive what is the minimal number of samples, $M_c(q)$, above which, for a given order of the moment (q), the bias of the logarithm of the sample moment estimator becomes approximately unbiased in the limit of large samples and large q.

The report is organized as follows: In chapter 2 we review previous studies of the convergence of averages of random variables. First, we describe an approach based on the consistency of the sample moment estimator. A different approach, developed in the area of statistical mechanics is described for the specific case of sums of iid lognormal random variables made in the random energy model [10]. Large Deviation Theory provides a formal mathematical framework for the results of the REM analysis. The application of the REM and LDT analysis to different types of distributions (Gaussian, Exponential, Chi-Squared and Exponential distributions) is made in chapter 3. LDT can only be applied when the distributions have non-algebraic decay. A separate analysis is required for Pareto distribution, because of the power-law form of its tail decay. Since sums of random variables appear in a great number of scientific areas, the results presented in this report can be applied in different contexts. The conclusions, results and perspectives of the investigation, together with possible applications in the areas of telecommunications, statistics and machine learning are presented in chapter 4. Detailed derivations of the results for the distributions considered are given in appendices A to D. Appendix E reviews the saddle point method for the asymptotic evaluation of integrals. This method is used repeatedly in the analysis developed. Finally, a generalization of Jensen's inequality, which is necessary to deduce the monotonicity of B_M, is proved in appendix F.

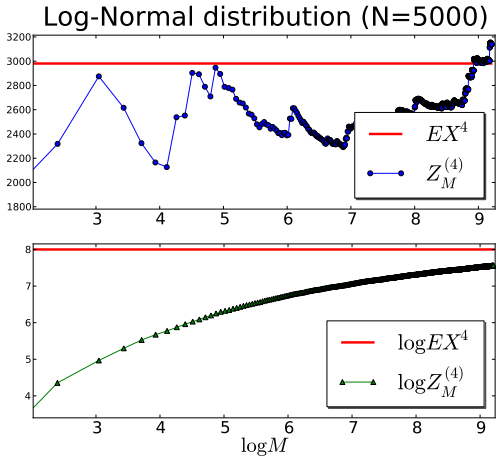


(a)

(b)



(c)



(d)

Figure 1.2: $Z_M^{(q)}$ and $\log Z_M^{(q)}$ for $q = 2, 4$ where each point is the average of N simulated Z_M with $\mu = 0$ and $\sigma = 1$.

Chapter 2

Previous work

As illustrated in the introduction the laws of large numbers and the Central Limit Theorem cannot be used to describe, even in an approximate manner, the behavior of averages in small samples. The purpose of this chapter is to review published studies that also analyze these limitations. Most of the investigations in the literature address the problem of how many moments can be estimated from a sample of finite size M. The goal is to find a critical value $q_c(M)$ such that lower moments $\mathbb{E}[X^q]$, whose order is below this critical value ($q < q_c(M)$) can be estimated with a certain accuracy, whereas higher-order moments ($q > q_c(M)$) cannot.

Kagan and Nagaev [12] define this critical value by requiring that the moment estimator be consistent for moments $q \leq q_c(M)$. The condition for consistency is that the probability of the expected value of the finite sample estimator being arbitrarily close to $\mathbb{E}\left[X^q\right]$ converges to one for $q \leq q_c(M)$.

This result they obtain is

$$q_c(M) = \frac{\log M}{2 \log \log M}$$

for distributions that satisfy the Bernstein's condition (see 2.11). For these types of distributions if $M = 10^6$, the critical order moment is $q_c(M = 10^6) = 2.63$. This result indicates that the requirement of consistency is probably too stringent to be useful in practical applications.

The statistical properties of $Z_M^{(q)}$ can be analyzed using a different approach. Specifically, one can use the techniques developed in statistical physics to investigate the properties of systems whose energy levels are random variables. The *random energy model* was first introduced by Derrida in [10] as a simplified model of disordered systems. The quantity of interest is the partition function of the system. Assume that $\{E_i\}_{i=1}^M$, the set of energy levels of the disordered system are iidrv's with density $f_E(e)$. In the original version of REM $f_E(e)$ is a normal distribution

$$f_E(e) = \frac{1}{\sqrt{\pi K}} \exp\left[-\frac{e^2}{K}\right], e \in (-\infty, \infty) \tag{2.1}$$

Nonetheless, the analysis is not restricted to this type of distribution. The partition function at temperature β^{-1} is defined as

$$Z_M(\beta) = \sum_{i=1}^{M} e^{-\beta E_i}. \tag{2.2}$$

From this partition function one can define a quenched free energy [13]

$$F_{M,q}(\beta) = -\frac{1}{\beta}\mathbb{E}\left[\log Z_M(\beta)\right] \tag{2.3}$$

The annealed free energy is

$$F_{M,a}(\beta) = -\frac{1}{\beta}\log\mathbb{E}\left[Z_M(\beta)\right] \tag{2.4}$$

Below a critical temperature ($\beta^{-1} < \beta_c^{-1}(M)$) the values of the quenched and the annealed free energy differ. As the temperature increases $F_{M,q}(\beta)$ approaches $\leq F_{M,a}(\beta)$ from below. Above certain threshold ($\beta_c^{-1}(M) < \beta^{-1}$) both free energies coincide. The system is said to be self-averaging. In the limit $M \to \infty$, this change of regime is a second-order phase transition in terms of the entropy function. For the standard REM, in which the energy levels are sampled from a normal distribution

$$s_a(y) = \log 2 - y^2 \tag{2.5}$$

It is a second order phase transition because the second derivative of the free energy is discontinuous at the transition temperature.

Derrida's model can be understood in terms of *Large Deviation Theory* (LDT), a branch of statistics that deals with the probability of large deviations in stochastic processes. In this theory, one assumes that the density function of the random variable can be approximated as the exponential of another function (an *entropy* in statistical physics, or a *rate function* in statistics) in the limit that a certain parameter (the system size, in statistical physics) is large. LDT provides the conditions under which the rate function (entropy) exists and the techniques to construct it explicitly.

As mentioned earlier, LDT can only be applied to certain types of distributions. For other cases, it is possible to use an approach based on *truncated moments*. The estimate of the qth moment in a finite sample $\{X_i\}_{i=1}^{M}$ should be close to the truncated estimate

$$Z_M^{(q)} \approx \int_0^{U(M)} x^q f_X(x)dx \tag{2.6}$$

where $U(M)$ is such as the probability of observing samples in $\{x_i\}$ greater than $U(M)$ is sufficiently small. It seems reasonable to choose $U(M)$ as a statistic of the maximum of the sample, such as the mean of the mode. In some cases, when closed-form expressions of these statistics are not available, one can approximate the distribution of the maximum using Extreme Value Theory [14]. In this case, $U(M)$ is chosen as the mean or the mode of the limit distribution of the sample maximum.

2.1 Convergence of moment estimators based on their consistency

In this section we review the investigation of A. Kagan and S. Nagaev [12]. These authors consider the problem of the simultaneous estimation of q moments from a sample of iidrv's of

size M. They determine a threshold $q_c(M)$, such that the sample estimator of the moment of order q becomes inconsistent for $q > q_c(M)$.

By the law of large numbers the sample average

$$Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q \tag{2.7}$$

is a strongly consistent estimator of $\mathbb{E}\left[X^q\right]$. This means that for any given $\epsilon > 0$

$$\mathbb{P}\left(|Z_M^{(q)} - \mathbb{E}\left[X^q\right]| > \epsilon\right) \underset{M\to\infty}{\longrightarrow} 0 \tag{2.8}$$

From this condition it follows that for any fixed integer $q > 0$ and $\epsilon > 0$

$$\mathbb{P}\left(\max_{1\le j\le q} |Z_M^{(j)} - \mathbb{E}\left[X^j\right]| > \epsilon\right) \le \sum_{j=1}^{q} \mathbb{P}\left(|Z_M^{(j)} - \mathbb{E}\left[X^j\right]| > \epsilon\right) \underset{M\to\infty}{\longrightarrow} 0 \tag{2.9}$$

Therefore, there must exist $q_c(M)$, dependent on the sample size M, such that

$$\mathbb{P}\left(\max_{1\le j\le q_c(M)} |Z_M^{(j)} - \mathbb{E}\left[X^j\right]| > \epsilon\right) \le \sum_{j=1}^{q_c(M)} \mathbb{P}\left(|Z_M^{(j)} - \mathbb{E}\left[X^j\right]| > \epsilon\right) \underset{M\to\infty}{\longrightarrow} 0 \tag{2.10}$$

and $\lim_{M\to\infty} q_c(M) = \infty$.

Consider distributions that satisfy Bernstein's conditions

$$|\mathbb{E}\left[X^q\right]| \le q! H^q,\ H \ge 0 \tag{2.11}$$

where H is a positive constant. This condition is equivalent to imposing analyticity of the characteristic function of the distribution in the interval $\left(-\frac{1}{H}, \frac{1}{H}\right)$. Some common distributions, such as the normal distribution, satisfy this condition. Other distributions, such as the lognormal, does not satisfy (2.11) because their characteristic function is not analytic at zero.

For these types of distributions, we can formulate the following theorem

Theorem 2.1.1 *Let $Z_M^{(j)}$ be the estimate of $\mathbb{E}\left[X^j\right]$ made in a sample of size M*

$$\lim_{M\to\infty} \mathbb{P}\left(\max_{1\le j\le q_c^-(M;\delta)} |Z_M^{(j)} - \mathbb{E}\left[X^j\right]| < \epsilon\right) = 1, \tag{2.12}$$

for any $\epsilon > 0$, where $q_c^-(M;\delta) = \dfrac{(1-\delta)\log M}{2\log\log M}(1 + o(1))$ for some $\delta \ge 0$.

This theorem establishes the consistency to order $q_c(M) \equiv q_c^-(M;\delta = 0)$. The following theorem describes the behavior of the sample estimator for orders $q > q_c(M)$.

Theorem 2.1.2 *In the same conditions as the previous theorem,*

$$\lim_{M\to\infty} \mathbb{P}\left(\max_{1\le j\le q_c^+(M;\delta)} |Z_M^{(j)} - \mathbb{E}\left[X^j\right]| > C \right) = 1 \tag{2.13}$$

for any $C > 0$, *and* $q_c^+(M;\delta) = \dfrac{(1+\delta)\log M}{2\log\log M}(1+o(1))$.

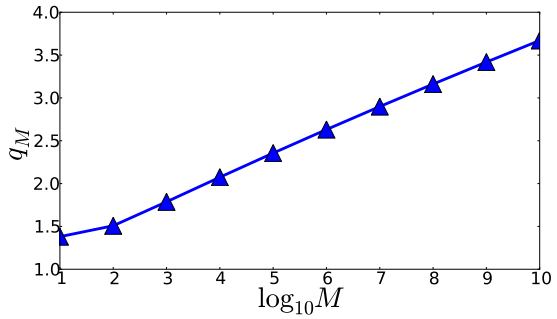
As a consequence of these two theorems,

$$q_c(M) = \frac{\log M}{2\log\log M} \tag{2.14}$$

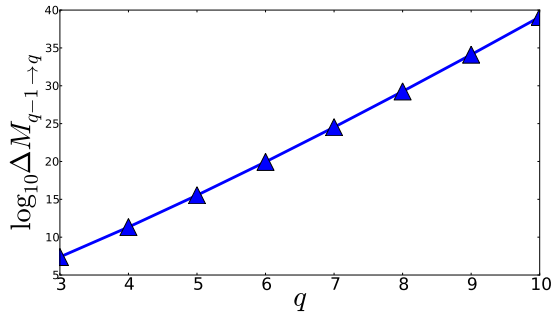
establishes that the maximum number of moments that can be estimated consistently from a sample of size M for $M \to \infty$. Even though

$$\lim_{M\to\infty} q_c(M) = \infty, \tag{2.15}$$

the growth of $q_c(M)$ with M is rather slow, as shown in Figure 2.1a.



(a)



(b)

Figure 2.1: (a) Order of moments that can be estimated as a function of the sample size. (b) Differences of the required sample size between $q-1$ and q.

For instance, with a sample of $M = 10^6$ the highest moment order that can be estimated is $q_c(10^6) = 2.63$. The slow growth of the critical order as the sample size increases seems

to be too restrictive to consider $q_c(M)$ a useful bound in practical applications. In the next section, we introduce the analysis of the random energy model, a model of disordered systems that exhibits a phase transition as a function of temperature in the appropriate scaling limit. The relevance of the analysis of REM to the current investigation is that it can be adapted to describe a phase transition in the sample estimator of moments of random variables.

2.2 The random energy model

The random energy model (REM) was introduced by Derrida in the area of statistical mechanics to describe the thermodynamic properties of disordered systems [10]. Consider a system with $M = 2^K$ configurations. The energy of the ith configuration is E_i. In the REM, we assume that the energy levels $\{E_i\}_{i=1}^M$ are iidrv's distributed following the pdf $f_E(e)$.

The partition function at temperature β^{-1} for a particular realization of the system is defined as

$$Z_M(\beta) = \sum_{i=1}^{M} e^{-\beta E_i}. \tag{2.16}$$

Thermodynamic quantities of the system such as average energy, energy fluctuations, heat capacity, etc. can be obtained from this quantity by taking derivatives with respect to the temperature [13]. Note that the partition function corresponds to the empirical estimate of the moment-generating function from the sample $\{E_i\}_{i=1}^M$.

For the subsequent derivations, it is useful to define the following relation.

Definition: Let A_M and B_M be two sequences of random variables

$$A_M \doteq B_M \Leftrightarrow \lim_{M\to\infty} \frac{1}{M} \log A_M = \lim_{M\to\infty} \frac{1}{M} \log B_M \tag{2.17}$$

This definition is also valid when the variables are indexed $K = \log M / \log 2$ and the limit $M \to \infty$ is taken.

Let $N(\epsilon, \epsilon+\delta)$ be the number of configurations in the interval $\mathcal{I} = [K\epsilon, K(\epsilon+\delta)]$. This quantity is a binomial random variable whose first two moments are

$$\begin{aligned} \mathbb{E}\left[N(\epsilon,\epsilon+\delta)\right] &= 2^K \mathcal{P}_{\mathcal{I}}\left(\epsilon,\epsilon+\delta\right) \\ Var\left[N(\epsilon,\epsilon+\delta)\right] &= 2^K \mathcal{P}_{\mathcal{I}}\left(\epsilon,\epsilon+\delta\right)\left(1-\mathcal{P}_{\mathcal{I}}\left(\epsilon,\epsilon+\delta\right)\right) \end{aligned} \tag{2.18}$$

where

$$\mathcal{P}_{\mathcal{I}}(\epsilon,\epsilon+\delta) = \int_{K\epsilon}^{K(\epsilon+\delta)} f_E(e)de = \int_{\epsilon}^{\epsilon+\delta} K \exp\left[K g_E(z)\right] dz \doteq \exp\left[K \max_{z\in[\epsilon,\epsilon+\delta]} \left\{g_E(z)\right\}\right]. \tag{2.19}$$

and $g_E(z)$ is a non-positive function. This last expression is obtained by saddle point integration (see appendix E).

The *microcanonical entropy density function* (in this work, to simplify the terminology, the *entropy function*) is defined as

$$\lim_{M\to\infty} \frac{1}{M} \log N(\epsilon, \epsilon+\delta) = \max_{y\in[\epsilon,\epsilon+\delta]} s_a(y) \tag{2.20}$$

$$N(\epsilon, \epsilon+\delta) \doteq \exp\left[K \max_{y\in[\epsilon,\epsilon+\delta]} s_a(y)\right] \tag{2.21}$$

$s_a(y)$ is named.

Using expression (2.18), in the leading exponential order,

$$s_a(y) = \log 2 - g_E(z) \tag{2.22}$$

The following proposition allows us to establish a relationship between the partition functions and the entropy functions.

Proposition 2.2.1 *If $s_a(y)$ exists and the limit in (2.20) is uniform then*

$$Z_M(\beta) \doteq \exp\left[K \max_y \{s_a(y) - \beta y\}\right]$$

In the original formulation of the random energy model [10], the energy levels are assumed to be normally distributed $\mathcal{N}(0, K/2)$

$$f_E(e) = \frac{1}{\sqrt{\pi K}} \exp\left[-\frac{e^2}{K}\right], e \in (-\infty, \infty) \tag{2.23}$$

For this particular model

$$\mathcal{P}_{\mathcal{I}}(\epsilon, \epsilon+\delta) = \int_{K\epsilon}^{K(\epsilon+\delta)} \frac{1}{\sqrt{\pi K}} \exp\left[-\frac{e^2}{K}\right] de = \sqrt{\frac{K}{\pi}} \int_{\epsilon}^{\epsilon+\delta} \exp\left[K(-y^2)\right] dy \tag{2.24}$$

Therefore, the entropy function is

$$s_a(y) = \log 2 - y^2 \tag{2.25}$$

Applying proposition (2.2.1),

$$Z_M(\beta) \doteq \exp\left[K \max_{y\in[y_l,y_u]} \left\{\log 2 - y^2 - \beta y\right\}\right] \tag{2.26}$$

because $s_a(y) \geq 0$, $y \in [y_l, y_u]$, where $y_l = -\sqrt{\log 2}$ and $y_u = \sqrt{\log 2}]$. For $\beta > 2\sqrt{\log 2}$ (low temperature), the exponent (2.26) does not have any local maximum in $[y_l, y_u]$. In this case, the maximum of the exponent is reached at y_l

$$Z_M(\beta) \doteq \exp\left[K\sqrt{\log 2}\beta\right] \tag{2.27}$$

For $\beta < 2\sqrt{\log 2}$ (high temperature), the exponent of (2.26) has a local maximum at $y^* = -\beta/2 \in [y_l, 0]$

$$Z_M(\beta) \doteq \exp\left[K\left(\log 2 + \frac{\beta^2}{4}\right)\right] \tag{2.28}$$

Therefore, there is a phase transition in $Z_M(\beta)$, as a function of β

$$\lim_{K\to\infty} \frac{1}{K} \log Z_M(\beta) = \begin{cases} \sqrt{\log 2}\beta, & \beta > \beta_c \quad \text{(low temperature)} \\ \log 2 + \frac{\beta^2}{4}, & \beta < \beta_c \quad \text{(high temperature)} \end{cases} \tag{2.29}$$

for $\beta_c = 2\sqrt{\log 2}$.

The transition is second order because there is a discontinuity in the second derivative of $Z_M(\beta)$ at the transition point $\beta_c = 2\sqrt{\log 2}$,

$$\frac{d}{d\beta}\left(\lim_{K\to\infty} \frac{1}{K} \log Z_M(\beta)\right)\Big|_{\beta=\beta_c} = \sqrt{\log 2} \tag{2.30}$$

$$\frac{d^2}{d^2\beta}\left(\lim_{K\to\infty} \frac{1}{K} \log Z_M(\beta)\right)\Big|_{\beta=\beta_c} = \begin{cases} 0, & \beta > \beta_c \\ \frac{1}{2}, & \beta < \beta_c \end{cases} \tag{2.31}$$

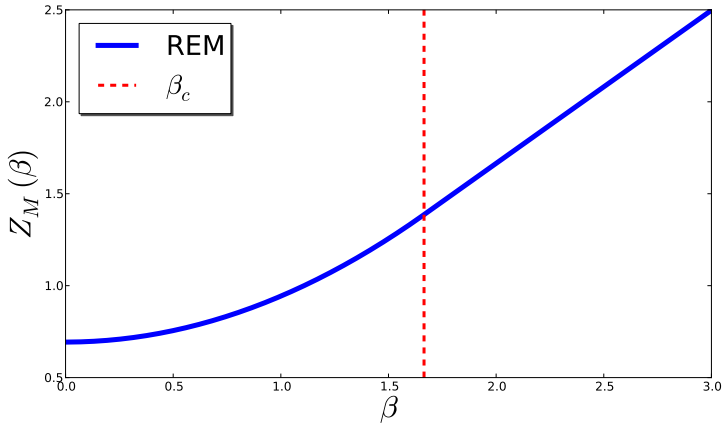


Figure 2.2: Illustration of the phase transition between the linear regime and the quadratic one.

The analysis made in the random energy model can be directly applied to the problem of estimating moments from a sample. Le X be a positive random variable whose pdf is $f_X(x)$. Consider the estimator of the qth moment from the sample of iidrv's $\{X_i\}_{i=1}^M$

$$Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q. \tag{2.32}$$

Define the random variable $Y = \log X$.

$$Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q = \frac{1}{M}\sum_{i=1}^{M} e^{qY_i} \tag{2.33}$$

where the pdf of the random variables Y_i is

$$f_Y(y) = f_X(e^y)e^y, y \in (-\infty, \infty) \tag{2.34}$$

The discontinuity that marks the phase transition for the random energy model appears only in the thermodynamic limit (the number of energy levels approaches infinity). For the moment estimation problem the scaling limit is taken in terms of the two parameters that appear in the problem: M, the sample size, and q, the order of the moment

$$M \to \infty,\ q \to \infty \ \text{ and } \ \frac{\log M}{h(q)} = 1 \tag{2.35}$$

where $h(q)$ is a monotonic increasing function of the order of the moment q.

2.2.1 Large Deviation Theory

Large Deviation Theory is a branch of statistics that studies the properties of the distributions of rare events, whose probability is assumed to decay exponentially with their size. This theory provides formal support to the analysis carried out in the Random Energy Model. In particular, it specifies the conditions under which the entropy function (in LDT terminology, the rate function) exists. The relationship between REM and LDT has been explored in [15, 16].

Definition: Let $\{Z_M\}$ be a sequence of random variables. The quantity $\mathbb{P}(Z_M \in A)$ satisfies a Large Deviation Principle with rate function $I_A(y)$ if the limit

$$-\lim_{M\to\infty} \frac{1}{M} \log \mathbb{P}(Z_M \in A) = I_A(y) \tag{2.36}$$

$$\mathbb{P}(Z_M \in A) = \exp\left[-MI_A + o(M)\right] \tag{2.37}$$

for large M.

In the case of discrete random variables, the rate function is defined as

$$-\lim_{M\to\infty} \frac{1}{M} \log P(Z_M = z) = I(z). \tag{2.38}$$

For continuous random variables, one needs to carry out an auxiliary step and define the density

$$\mathbb{P}\left(Z_M \in [z, z+dz]\right) = f_{Z_M}(z)dz \tag{2.39}$$

in the limit $dz \to 0$ and $M \to \infty$. By the definition of the rate function (2.36),

$$-\lim_{M\to\infty} \frac{1}{M} \log f_{Z_M}(z) - \lim_{M\to\infty} \frac{1}{M} \log dz = I(z).$$

Therefore,

$$f_{Z_M}(z) = \exp\left[-MI(z) + o(M)\right] \tag{2.40}$$

where the second term is an arbitrary non-zero infinitesimal value that can be neglected in the exponential leading order.

In the limit $M \to \infty$ the expected values of $g(Z_M)$

$$\mathbb{E}\left[g(Z_M)\right] = \int g(z) f_{Z_M}(z) dz \doteq \int g(z) \exp(-MI(z)) dy \tag{2.41}$$

can be approximated asymptotically using saddle point integration techniques (appendix E, [17]).

However, ensuring the existence of a rate function is not a trivial question. To formulate these conditions, we need the following definition:

Definition: Given a sequence of random variables $\{Z_M\}$ we define the *scaled cumulant generating function* as the limit

$$\lambda(k) = \lim_{M\to\infty} \frac{1}{M} \log \mathbb{E}\left[\exp\left[kMZ_M\right]\right] \tag{2.42}$$

Next, we formulate the Gärtner-Ellis theorem. This theorem specifies the conditions under which the probability distribution can be expressed in terms of a rate function

Theorem 2.2.2 *If $\lambda(k)$ exists and is differentiable $\forall k \in \mathbb{R}$, then Z_M satisfies a Large Deviation Principle. The rate function is*

$$I(z) = \mathcal{L}(\lambda) = \sup_{k\in\mathbb{R}} \{kz - \lambda(k)\} \tag{2.43}$$

The functional applied on $\lambda(k)$ is called the Legendre-Fenchel transform. The formula for the rate function $I(z)$ can be derived from $\lambda(k)$ by first approximating $\langle \exp\left[kMZ_M\right]\rangle$ as $M \to \infty$ using saddle point estimation and then noting Legendre-Fenchel transform is self-inverse, that is, $\mathcal{L}\left(\mathcal{L}(\lambda(k))\right) = \lambda(k)$.

For the sample average $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$ the computation of the rate function is particularly simple

$$\lambda(k) = \lim_{M\to\infty} \frac{1}{M} \mathbb{E}\left[\exp\left[kMZ_M\right]\right] = \lim_{M\to\infty} \frac{1}{M} \log \mathbb{E}\left[\prod_{i=1}^{M} \exp\left[kX_i\right]\right] = \tag{2.44}$$

$$= \lim_{M\to\infty} \frac{1}{M} \log \prod_{i=1}^{M} \mathbb{E}\left[\exp\left[kX_i\right]\right] = \log \mathbb{E}\left[\exp\left[kX\right]\right] \tag{2.45}$$

Therefore, to compute $\lambda(k)$ for sums of random variables, one only needs to calculate the cumulant generating function. The differentiability condition in the Gärtner-Ellis theorem is satisfied because the Laplace transform (i.e. the moment generating function) is always differentiable when it is defined.

Example: Let X be a normal random variable of mean μ and variance σ^2. Let $Z_M = \frac{1}{M}\sum_{i=1}^{M} X_i$,

$$\lambda(k) = \mu k + \frac{1}{2}\sigma^2 k^2 \tag{2.46}$$

$\lambda(k)$ is infinitely differentiable therefore we can compute the rate function by applying the Gärtner-Ellis theorem and we obtain $I(z) = \frac{(z-\mu)^2}{2\sigma^2}$.

Since the Laplace transform is always differentiable where it is defined, we can apply the Gärtner-Ellis theorem and compute the rate function for distributions whose cumulant generating function is only partially defined; that is, if it is not defined $\forall k \in \mathbb{R}$.

Example: Let X be a exponential random variable with scale parameter 1,

$$\lambda(k) = -\log(1-k) \tag{2.47}$$

and by just computing the maximum in (2.43), $I(z) = z - 1 - \log z$.

Important cases, such as the power-law distribution, are out of the scope of Large Deviation Theory.

Example: Let X be a Pareto random variable. Its moment generating function is

$$M(k) = \mathbb{E}[e^{kX}] = \int_1^\infty \alpha x^{-\alpha-1} e^{kx} dx \tag{2.48}$$

Note $M(k) < \infty$ only if $k \leq 0$. In this case

$$\lambda(k) < \lambda(k+\delta) \quad \forall \delta > 0 \tag{2.49}$$

because $e^{kx} < e^{(k+\delta)x}$ and $\alpha x^{-1-\alpha}$ is positive in $[1, \infty)$. Using this observation and $M(0) = 1$, the maximum in (2.43) is reached at $k = 0$,

$$I(y) = 0 \quad \forall y \in [1, \infty) \tag{2.50}$$

Therefore, for distributions with an algebraic tail decay, zero rate functions are obtained. The reason is that the tails of the distribution decays slower than any exponential function of M [15].

Finally, we state Varadhan's Theorem, which is a generalization of the Gärtner-Ellis Theorem. Given a continuous function g, we define

$$\lambda(g) = \lim_{M\to\infty} \frac{1}{M} \log \mathbb{E}\left[\exp\left[Mg(Z_M)\right]\right]. \tag{2.51}$$

The rate function $I(y)$ is obtained by taking the Legendre-Transform of g,

$$\lambda(g) = \sup_y \{g(y) - I(y)\} \tag{2.52}$$

This theorem will be useful in the derivation of the random energy model results from LDT.

We now detail how the random energy model can be formulated using large deviation theory [18, 16]. The partition function of the random energy model is

$$Z_M(\beta) = \sum_{i=1}^{M} \exp\left[-\beta E_i\right], \quad M = 2^K, \tag{2.53}$$

where E_i are Gaussian random variables whose pdf is

$$f_E(e) = \frac{1}{\sqrt{\pi K}} \exp\left[-\frac{e^2}{K}\right] \tag{2.54}$$

Given a random sample $\{E_i\}$, we can construct a probability distribution F_K as

$$F_K(x) = \frac{|\left\{i : \frac{E_i}{K} \leq x\right\}|}{2^K} \quad i = 1, \ldots, 2^K. \tag{2.55}$$

Note that F_K is a discontinuous distribution defined on $\mathbb{R}$. It has jumps at the points $\{E_i\}$. If X is the random variable with cdf $F_K(x)$

$$\mathbb{P}\left(X \in \left[\frac{E_i}{K}, \frac{E_{i+1}}{K}\right)\right) = 0 \tag{2.56}$$

$$\mathbb{P}\left(X \in \left[\frac{E_i}{K}, \frac{E_{i+1}}{K}\right]\right) = \frac{1}{2^K} \tag{2.57}$$

Taking the expected value of $h(x) = e^{-\beta K x}$, we obtain

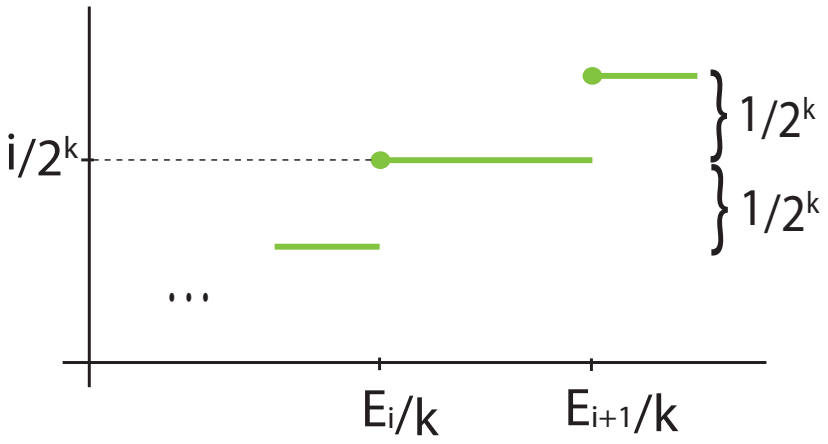
$$\begin{aligned} E\left[h(x)\right] &= \int_{-\infty}^{\infty} h(x) dF_K(x) = \sum_{i=1}^{2^K} h\left(\frac{E_i}{K}\right) \frac{1}{2^K} \\ &= \frac{1}{2^K} \sum_{i=1}^{M} e^{-\beta E_i} = \frac{1}{2^K} Z_M(\beta) \end{aligned} \tag{2.58}$$

Therefore, Z_M can be expressed in terms of the expected value with respect to a a probability distribution that depends on K

$$Z_M(\beta) = 2^K \int_{-\infty}^{\infty} e^{-\beta K x} dF_K(x) = 2^K \mathbb{E}\left[e^{-\beta K x}\right] \tag{2.59}$$

Theorem 2.2.3 *With probability 1, the measure induced by F_K satisfies a Large Deviation Principle with index K and rate function*

$$I(z) = \begin{cases} z^2, \text{if } |z| \leq \sqrt{2 \log 2} \\ \infty, \text{if } |z| > \sqrt{2 \log 2} \end{cases} \tag{2.60}$$



(a)

Figure 2.3: Measure of probability dF_K

From the definition of $\lambda(g)$ and taking $g(z) = -\beta z$,

$$\lambda(g) = \lim_{K\to\infty} \frac{1}{K} \log \mathbb{E}\left[\exp\left[K(-\beta X)\right]\right] \tag{2.61}$$

Applying Varadhan's Theorem,

$$\lim_{K\to\infty} \frac{1}{K} \log \mathbb{E}\left[\exp\left[K(-\beta X)\right]\right] = \sup_z \left\{-z^2 - \beta z\right\} \tag{2.62}$$

Finally,

$$\lim_{K\to\infty} \frac{1}{K} \log Z_M = \sup_y \left\{\log 2 - x^2 - \beta x\right\} \tag{2.63}$$

which is the result derived earlier (2.33). In the next sections the methods developed to analyze the random energy model are applied to the problem of estimation of moments from a finite sample.

Chapter 3

Convergence of Monte Carlo averages

In the previous section, we studied how the analysis of the partition function in the Random Energy Model reveals the existence of a phase transition as a function of the system temperature. In this chapter we extend and adapt this analysis to the estimation of the moment of order q from a sample of finite size $\{X_i\}_{i=1}^M$

$$Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q = \frac{1}{M}\sum_{i=1}^{M} e^{qY_i}, \tag{3.1}$$

where the change of variables $X_i = e^{Y_i}$ has been made in the last step. Note this way of writing the estimator $Z_M^{(q)}$ brings out the similarity between a sample moment estimator and the REM partition function in 2.16. As argued in the previous sections, the convergence of $Z_M^{(q)}$ to its infinite sample limit $\mathbb{E}[X^q]$ can be analyzed in terms of the bias of the *logarithm* of (3.1)

$$\mathbb{E}[B_M^{(q)}] = \log \mathbb{E}[X^q] - \mathbb{E}[\log Z_M^{(q)}]. \tag{3.2}$$

Formally, if one chooses $q = -\beta$ and $Y_i \sim \mathcal{N}(0, K/2)$ where $K = \log M / \log 2$, the moment estimator (3.1) becomes the partition function of the random energy model. However, in the general context of moment estimation, the parameters of the distribution are fixed and do not depend on the sample size. Nonetheless, it is possible to make a derivation that is parallel to the analysis of the random energy model and define an entropy function, which could depend in a complicated manner on K. Even though, in this case, it is sometimes not possible to derive closed-form expressions as in the standard REM, the interval $[y_l, y_u]$ in which the entropy function is positive can always be estimated numerically. Once this interval has been computed, one can carry out a derivation similar to REM and find an exponential approximation to the sample estimate of the qth moment

$$Z_M^{(q)} \doteq \frac{1}{M} \exp\left[K \max_{y\in[y_l,y_u]} \{s_a(y) + qy\}\right] \tag{3.3}$$

For a fixed value of q, depending on the sample size, the maximum of the exponent could be either inside the interval or at y_u:

1. If $M < M_c(q)$, the main contribution to (3.3) comes from the upper bound of the interval y_u. In this case $\mathbb{E}[B_M^{(q)}] > 0$ and the typical values of $Z_M^{(q)}$ are a poor estimation of $\mathbb{E}[X^q]$.

2. If $M_c(q) < M$, the dominance corresponds to y^*, the local maximum of the exponent. In this case $Z_M^{(q)}$ is well approximated by the asymptotic form obtained by saddle point estimation. In this asymptotic limit the bias of the logarithm of the estimator disappears $\mathbb{E}[B_M^{(q)}] = 0$, which means that typical values of $Z_M^{(q)}$ are a good estimation of $\mathbb{E}[X^q]$.

In the limit of large q, there is an abrupt change of behavior in the estimator as a function of the sample size.

In this chapter this analysis is carried out for different types of distributions. In section 3.1 we consider $X \sim \log -\mathcal{N}(0, \sigma)$, that is, the random variable Y is Gaussian with zero mean and variance σ^2. As in the REM case, the interval of definition for the entropy function can be explicitly computed and, following the same reasonings, the phase transition in the estimator can be described with expressions in closed form.

A generalization of the lognormal case is constructed in section 3.6 for exponential-power-law distributions [19]. The family of distributions considered in this section is a generalization of the lognormal case because the pdf of the random variable Y is

$$f_Y(y) = \rho L'(y) y^{\rho-1} \exp\left[-L(y) y^{\rho}\right] \tag{3.4}$$

where $L(y)$ is a slow varying function; that is, it behaves asymptotically as if it were a constant. The lognormal case is retrieved with $\rho = 2$ (see [19]). We also compare the exponential-power-law approach to the results obtained in section 3.1.

In sections 3.3, 3.4 and 3.5 we analyze the Normal, Weibull and Chi-Squared cases, respectively. It is interesting to analyze the Gaussian case because of the central role played by the Gaussian distribution in the convergence of sums of independent random variables, as stated in the Central Limit Theorem. On the other hand, the Weibull distribution is the distribution of the powers of exponential random variables. Finally, the Chi-Squared distribution with d degrees of freedom is related to Normal distribution: sum of d squared normal random variables is distributed as a Chi-Squared random variable. For these three cases explicit expressions for the bounds of the interval in which the entropy function is positive ($[y_l, y_u]$) cannot be found. The reason is that the equation for the zeros of the entropy function

$$s_a(y) = \log 2 - g_a(y) + \xi y, \tag{3.5}$$

where $g_a(y)$ is an exponential function of y, cannot be solved explicitly. Nonetheless, approximate solutions of (3.5) can be obtained using numerical methods. Alternatively, approximate closed-form solutions can be derived by balancing the two dominant terms in (3.5),

$$\log 2 + \xi \widehat{y}_l = 0 \Rightarrow \widehat{y}_l = -\log 2/\xi \tag{3.6}$$

$$\log 2 - g_a(\widehat{y}_u) = 0 \Rightarrow \widehat{y}_u = g_a^{-1}(\log 2) \tag{3.7}$$

By applying equation (3.3), as in the explicit lognormal case, using the approximated $\widehat{y}_l$ and $\widehat{y}_u$ closed-form expressions of the bias (3.2) on both sides of phase transition are obtained.

As discussed in section 2.2.1 an analysis similar to REM, or more generally, based on large deviation theory, cannot be applied to distributions with algebraic tails. The reason is that the entropy function is linear. Therefore, the maximum in (3.3) always corresponds to the lower bound y_l. To model $B_M^{(q)}$ in the Pareto distribution, we propose to use a method based on approximating the sample estimate of the qth moment by the *truncated moment*

$$Z_M^{(q)} \approx \int_0^{U_M} x^q f_X(x) dx \tag{3.8}$$

where U_M is such that $\mathbb{P}(\max_i \{X_i\}_{i=1}^M > U_M)$ is sufficiently small. It seems reasonable to make the choice in terms of the maximum of the sequence of random variables, $G_M = \max_i \{X_i\}_{i=1}^M$. The probability density function (pdf) of this maximum is

$$f_M(x) = M f(x) \left[F(x)\right]^{M-1} \tag{3.9}$$

where $F(x)$ is the probability distribution of X and $f(x)$ the corresponding density.

In the limit of large samples, Extreme Value Theory [14] provides accurate approximations to the distribution of G_M in terms of three different types of distributions. Specifically, the Extreme Value Theorem states that in the $M \to \infty$ the distribution of G_M approaches one of these three functional forms: (i) the Weibull distribution, (ii) the Gumbel distribution and (iii) the Fréchet distribution (see theorem 3.7.1). In section 3.7 the sample average of Pareto iid random variables are considered. EVT states that the asymptotic behaviour of G_M is of the Fréchet from when X is a Pareto random variable. We compare three models depending how the value of U_M is chosen: (a) the mode of the exact distribution of G_M, (b) the mode of the Fréchet distribution and (c) the mean of the Fréchet distribution.

3.1 The lognormal distribution

Since the partition function in the Random Energy Model (2.16) is, basically, the sum of lognormal random variables whose variance depends on the number of energy levels, it is natural to consider the estimator of the moments of lognormal random variables with fixed parameters.

A random variable X is said to be lognormal random variable of parameters (μ, σ) if $Y = \log X$ so that Y is a Normal random variable $\mathcal{N}(\mu, \sigma^2)$.

The parameter μ only appears in the scale factor e^{μ}. Therefore, one can assume $\mu = 0$ without loss of generality. Furthermore, if X is lognormal with parameters $(\mu = 0, \sigma)$, $X^q = \exp\left[q\sigma Y\right]$ is also lognormal with parameters $(\mu = 0, \sigma q)$. For this reason, it is sufficient to analyze the estimator of the moment of a lognormal with parameters $\log - \mathcal{N}(\mu = 0, \sigma)$,

$$Z_M^{(\sigma)} = \frac{1}{M} \sum_{i=1}^{M} X_i = \frac{1}{M} \sum_{i=1}^{M} \exp\left[Y_i\right], Y_i \sim \mathcal{N}(0, \sigma^2) \tag{3.10}$$

as a function of σ and M.

The framework developed for the random energy model can also be applied to the analysis of the sample average (3.10), provided that we make the appropriate parameter identifications. However, in contrast to REM, the parameters of the corresponding normal random variable are independent of M. Therefore one has to consider an appropriate asymptotic limit, which is defined as $\sigma \to \infty$ and $M \to \infty$ with $\log M/\sigma^2 \to$ constant. To carry out the analysis in this limit, one first computes the probability of finding samples in the interval $[K\epsilon, K(\epsilon + \delta)]$ to the exponential leading order. Using (2.22), the entropy function is

$$s_a(y) = \log 2 - \frac{Ky^2}{2\sigma^2} \tag{3.11}$$

Since, by definition, the entropy function $s_a(y)$ is non negative, the range of y is restricted to the interval $|y| \leq \sqrt{\frac{2\sigma^2 \log 2}{K}}$. Depending on the location of the maximum of the exponent in (2.2.1) we can distinguish two regimes: The first one appears for small samples $M < M_c(\sigma)$, where $M_c(\sigma) = e^{\sigma^2/2}$. Asymptotically, the dominant contribution to the estimator (3.10) comes from the upper bound of the interval $y_u = \sqrt{\frac{2\sigma^2 \log 2}{K}}$. The second regime corresponds to large samples $M > M_c(\sigma)$, in which the dominant contribution comes from the local maximum of the exponent $y^* = \dfrac{\sigma^2}{K}$. In this regime, and in the limit of large M and σ, the sample average is close to the saddle point estimate of the moment.

The change of behavior in $y^* = y_u$ corresponds to a second order phase transition in the asymptotic behavior of $Z_M^{(\sigma)}$

$$Z_M^{(\sigma)} \doteq \begin{cases} \exp\left[\frac{\sigma^2}{2}\right], \lambda(M,\sigma) > 1 \\ \\ \exp\left[\sqrt{2\sigma^2 \log M} - \log M\right], \lambda(M,\sigma) < 1 \end{cases} \tag{3.12}$$

where $\lambda(M,\sigma) = \dfrac{2\log M}{\sigma^2}$. The transition is marked by a discontinuity in the second derivative of $\mathbb{E}[\log Z_M]$ at the transition point in the limit $\sigma \to \infty$ and $M \to \infty$ with $\log M/\sigma^2 \to$ constant. There is no discontinuity for finite M and σ. Nevertheless, the approximation (3.12) is still accurate for sufficiently large values of these parameters. In particular, for a sample of size M, there is a critical value $\sigma_c(M) = \sqrt{2\log M}$, such that, for values of $\sigma > \sigma_c(M)$ the sample average ceases to be an accurate approximation of the expected value of the lognormal random variable. Alternatively for a fixed value σ, the critical sample size is $M_c(\sigma) = e^{\frac{\sigma^2}{2}}$.

To validate this analysis, we compare the asymptotic formulas for with simulations in terms of $B_M^{(\sigma)}/B_{max}^{(\sigma)}$.

Note also that an exact expression can be obtained for $B_{max}^{(\sigma)}$,

$$B_{max}^{(\sigma)} = \mathbb{E}[B_1] = \log \mathbb{E}[X] - \mathbb{E}[\log X] = \frac{1}{2}\sigma^2 \tag{3.13}$$

Table 3.1: Critical sample size M_c for some values of σ

Shape parameter σ	Critical size (M_c)
2	7.38
4	2.98×10^3
8	7.89×10^{13}
16	3.88×10^{55}

since $\mathbb{E}[X] = \exp\left[\frac{1}{2}\sigma^2\right]$ and $\mathbb{E}[\log X] = 0$ by observing $\log X$ is a Normal random variable with zero mean and standard deviation σ. Using equation (3.13) and by definition of B_M (3.2),

$$\widehat{B}_M^{(\sigma)} = \frac{B_M^{(\sigma)}}{B_{max}^{(\sigma)}} = \begin{cases} 0, & \lambda > 1 \\ \dfrac{2\left(\sqrt{2\sigma^2 \log M} - \log M\right)}{\sigma^2}, & \lambda < 1 \end{cases} \tag{3.14}$$

In table 3.1 some values of M_c as a function are showed. It can be observed that critical sample sizes are smaller than those predicted by Kagan and Nagaev (see section 2.1).

As we can observe in figure 3.1, the model is more accurate when σ increases. Even though it is not necessary to have large values of σ to obtain accurate models, it can be observed that the theoretical transition point is far from the empirical one for low values of σ.

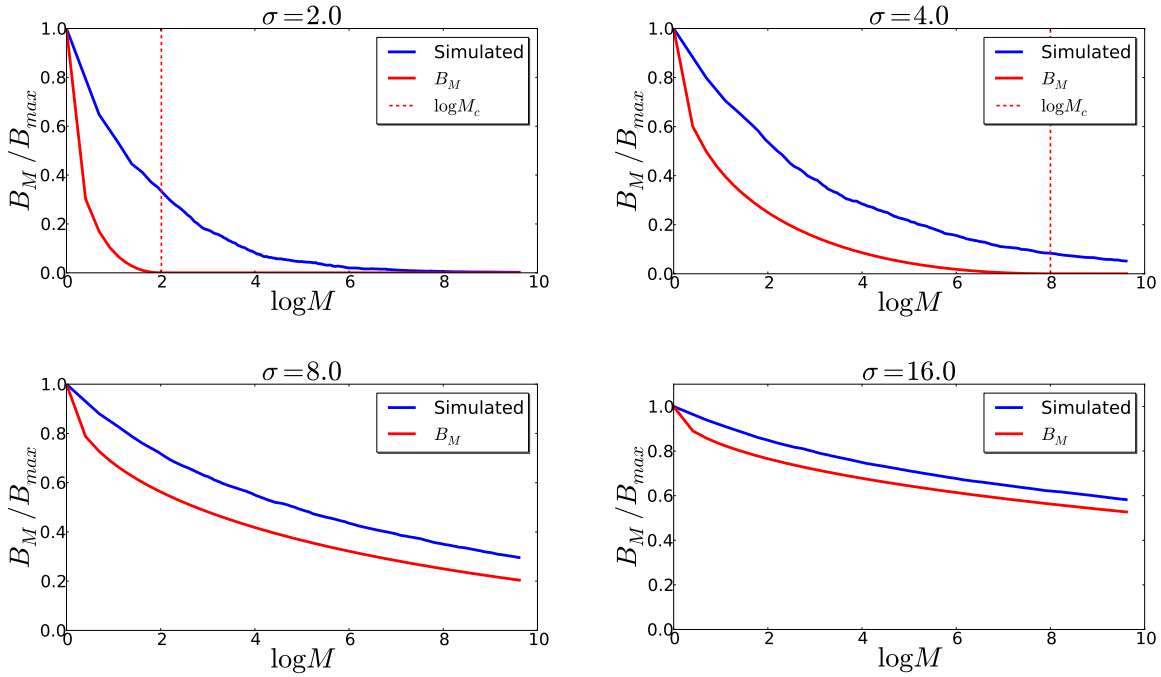


Figure 3.1: Comparison of the theoretical model to simulations for lognormal distribution.

A generalization of the results of the Random Energy Model has been developed in order to study the estimators of the moments of lognormal distributions. As well as in the REM a

phase transition was obtained in the behaviour of the partition function (2.29), we have also obtained a phase transition in the limit $M \to \infty$, $\sigma \to \infty$ in the estimator $Z_M^{(\sigma)}$. In the next sections, we generalize this analysis to different distributions.

3.2 The Log-Exponential-Power-Law distribution

In this section we deal with the analysis of $Z_M^{(q)}$ by reviewing the study of the family of Log-Exponential-Power-Law distributions presented in [19]. The results of this section can be compared to those of the previous section because the lognormal distribution is a particular case of a Log-Exponential-Power-Law distribution.

We say X is a Log-Exponential-Power-Law random variable if the cdf of $Y = \log X$ satisfies

$$1 - F_Y(y) = \exp\left[-h(y)\right] \tag{3.15}$$

where the function $h(y) = L(y)y^\rho$, with $\rho > 1$ and d $L(y)$ is a slowly varying function, i.e.

$$\lim_{y\to\infty} \frac{L(ty)}{L(y)} = 1, \forall t \in \mathbb{R}.$$

We also impose a condition on the derivatives of $h(y)$,

$$\lim_{y\to\infty} \frac{h''(y)}{h'(y)} = 0 \tag{3.16}$$

Under these assumptions, the pdf of the random variable Y is given by

$$f_Y(y) = h'(y)\exp[-h(y)]$$

and the qth moment can be written as

$$\mathbb{E}[X^q] = \mathbb{E}[e^{yq}] = \int_{-\infty}^{\infty} \exp\left[qy - y^\rho L(y) + \log h'(y)\right] dy \tag{3.17}$$

Using the saddle point method to approximate quadratures (see appendix E)

$$E[X^q] \doteq \exp\left[qy^* - y^{*\rho} L(y^*) + \log h'(y^*)\right] \tag{3.18}$$

where y^* is the maximum of the argument of the exponential in the integrand. By the condition of local maximum and using (3.16)

$$q - h'(y^*) + \frac{h''(y^*)}{h'(y^*)} = 0 \Rightarrow h'(y^*) \approx q \tag{3.19}$$

We can approximate y^* by neglecting the slowly varying function $L(y)$,

$$y^* \approx \left(\frac{q}{\rho}\right)^{1/(\rho-1)} \tag{3.20}$$

This result provides a way to approximate y^* using only the derivative of $h(y)$.

Since this result is valid only asymptotically, when the sample size is large, the next step is to analyze the finite size effects in the moment estimate. The probability of the maximum of a sequence of random variables is given by

$$\mathbb{P}\left(\max_i \{Y_i\}_{i=1}^M < y\right) = \mathbb{P}\left(\{Y_i\}_{i=1}^M < y,\ \forall i\right) = \mathbb{P}(Y < y)^M$$

because of independence. Consider now a *threshold* y_τ^+ such that the probability of observing values that are larger than this threshold is sufficiently small. This quantity can be defined in terms of the distribution of the maximum,

$$\mathbb{P}\left(\max_i \{Y_i\}_{i=1}^M < y_\tau^+\right) = e^{-\tau} \tag{3.21}$$

Then,

$$1 - \mathbb{P}\left(\max \{Y_i\}_{i=1}^M < y_\tau^+\right) = 1 - \mathbb{P}\left(Y < y_\tau^+\right)^M = 1 - e^{-\tau}$$

The complementary of the cdf evaluated at the threshold is

$$1 - F_Y(y_\tau^+) = 1 - \exp\left[-\frac{\tau}{M}\right] = \frac{\tau}{M} + o\left(\frac{1}{M}\right),$$

where the exponential has been approximated using its Taylor series truncated to first order. By definition of the class of random variables we are considering,

$$h(y_\tau^+) = \log M - \log \tau + o\left(1\right)$$

Noting $h(y) \sim y^\rho$ and $h(y_1^+) \approx \log M$, by taking the quotient between y_τ^+ and y_1^+,

$$\begin{aligned} y_\tau^+ &\sim \left(1 + \frac{\log \tau}{\log M}\right)^{1/\rho} y_1^+ \\ h(y_\tau^+) &\sim \left(1 + \frac{\log \tau}{\log M}\right) \log M \approx \log M + \mathcal{O}(\tau) \end{aligned}$$

Therefore $h(y_\tau^+)$ behaves like $\log M$ with a small correction that tends to zero with $\log M$. Then, it is natural to require

$$h(y^+(M)) = \log M \tag{3.22}$$

Note that y_τ^+ becomes closer to y_1^+ as the sample size increases. This fact allows us to remove the dependency in τ since for any τ $\mathcal{O}(1)$, y_τ^+ is only a small correction of y_1^+. Proceeding as in equation (3.20),

$$y^+(M) \approx (\log M)^{1/\rho} \tag{3.23}$$

For a fixed q and as a function of sample size, the boundary between the two regimes is defined in terms of a critical sample size $M_c(q,\rho)$

$$y^+(M_c(q,\rho)) = y^*. \tag{3.24}$$

Using approximations from (3.20) and (3.23), the phase transition occurs in the limit $M \to \infty$, $q \to \infty$ and

$$\lambda(M,q,\rho) = \rho^{\rho/(\rho-1)} \frac{\log M}{q^{\rho/(\rho-1)}} \longrightarrow \text{constant} \tag{3.25}$$

When the sample is larger than $M_c(q,\rho)$, the empirical moment estimator is dominated by (3.18). For small samples we can use the estimation based on *truncated moments*

$$\begin{aligned} T_M^{(q)} &= \int_{-\infty}^{y^+(M)} \exp\left[qy - h(y) + \log h'(y)\right] dy \doteq \\ &\doteq \exp\left[qy^+(M) - h(y^+(M)) + \log h'(y^+(M))\right] \end{aligned} \tag{3.26}$$

The convergence to leading exponential order of $T_M^{(q)}$ to $Z_M^{(q)}$ is proved in [19].

In summary, for a fixed q, the sample average has a sharp change of behavior

$$Z_M \doteq \begin{cases} \exp\left[qy^+(M) - h(y^+(M)) + \log h'(y^+(M))\right], & M < M_c(q,\rho) \\ \exp\left[qy^*(M) - h(y^*(M)) + \log h'(y^*(M))\right], & M \geq M_c(q,\rho) \end{cases} \tag{3.27}$$

This change of behavior leads to a discontinuity in the second derivative of $\mathbb{E}[Z_M]$ in the limit $M \to \infty$, $q \to \infty$ and $\dfrac{\log M}{q^{\rho/(\rho-1)}} \to$ constant.

As mentioned earlier, the lognormal distribution is a particular case of the Log-Exponential-Power-Law distribution with $\rho = 2$. In the lognormal case, Y is a Gaussian random variable whose cdf satisfies

$$\begin{aligned} F_Y(y) &= \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{y}{\sqrt{2}}\right] \\ \exp\left[-h(y)\right] &= 1 - F_Y(Y) = \frac{1}{2} - \frac{1}{2}\text{erf}\left[\frac{y}{\sqrt{2}}\right] = \frac{1}{2}\text{erfc}\left[\frac{y}{\sqrt{2}}\right] \\ h(y) &= \log 2 - \log \text{erfc}\left[\frac{y}{\sqrt{2}}\right] \end{aligned} \tag{3.28}$$

where erf is the error function and erfc the complementary error function. Using the asymptotic approximation for the complementary error function

$$\text{erfc}(y) \underset{y\to\infty}{\longrightarrow} \frac{\exp(-y^2)}{\sqrt{2\pi}y} \tag{3.29}$$

and by (3.28) becomes,

$$h(y) \sim \log 2 + \frac{y^2}{2} + \log(\sqrt{\pi}y) \tag{3.30}$$

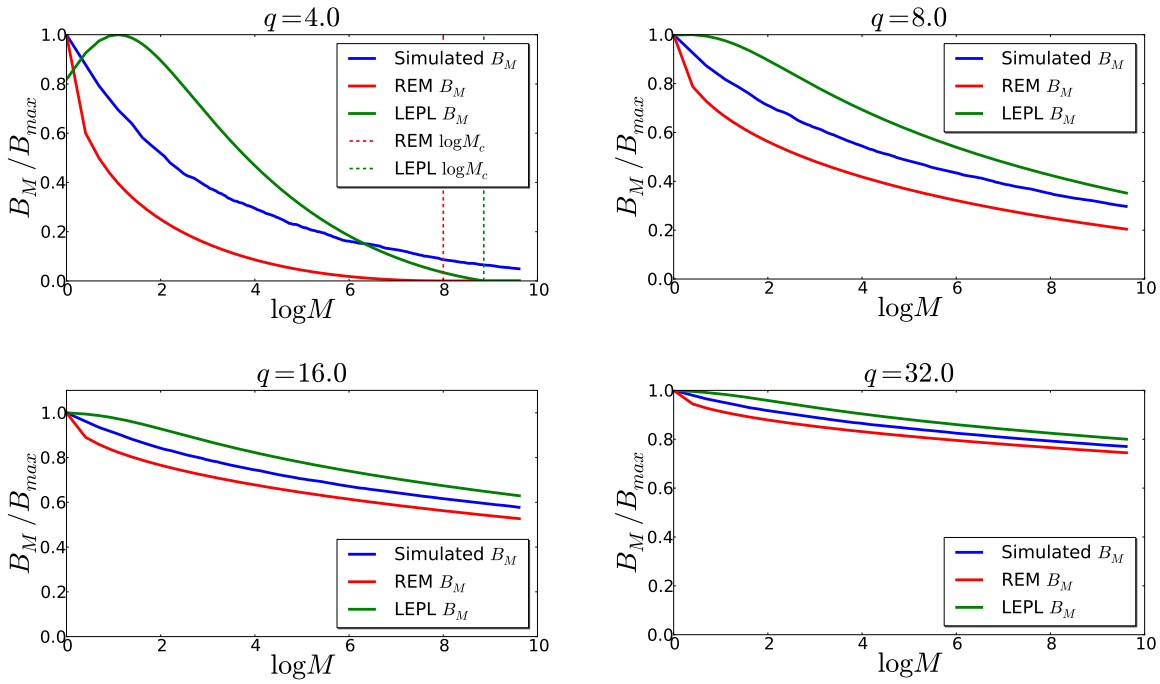


Figure 3.2: Comparison of the Log-Exponential-Power-Law model and the REM to simulations.

We use this approximation to $h(y)$ in conditions (3.19) and (3.22) to define y^* and y^+, respectively. In figure 3.2 we compare this model to the one in 3.1 and simulations.

We observe in figure 3.2 the approach of this section (green curve) achieves similar accuracy to section 3.1 model (red curve). A significant difference between both models is LEPL model overestimates the normalized bias where as REM approach underestimates it.

3.3 The Folded-Normal distribution

The analysis of the convergence of the empirical moments of the normal distribution is of particular interest because of the central role this distribution plays in statistics. In this section, we carry out an analysis parallel to section 3.1 to study the behaviour of the estimator $Z_M^{(q)}$.

Since the analysis presented is valid only for non-negative random variables, we consider the problem of estimating the moments of a folded normal random variables $X = |Z|,\ Z \sim \mathcal{N}(0,1)$ from a sample of size M. Note that the even moments of X coincide with those of a standard normal variable. From X, we define the random variable $Y = \log X$, whose pdf is

$$f_Y(y) = \frac{2}{\sqrt{2\pi}} \exp\left[-\frac{e^{2y}}{2} + y\right], y \in (-\infty, \infty) \tag{3.31}$$

Following (2.22), the entropy function $s_a(y)$,

$$s_a(y) = \log 2 - \frac{e^{2Ky}}{2K} + y \tag{3.32}$$

with $K = \log 2/\log M$. The entropy function is non-negative by definition. Therefore, it is non-zero only in a bounded interval $[y_l, y_u]$. In contrast to the lognormal and the log-exponential-power-law cases, no closed-form expressions can be given for the bounds of this interval. Nonetheless, numerical algorithms for finding zeros of nonlinear functions can be used to estimate them. Alternatively, approximate closed-form expressions can be obtained by balancing the two dominant terms in the equation $s_a(y) = 0$,

$$\begin{aligned} &\log 2 + \hat{y}_l = 0 \Rightarrow \hat{y}_l = -\log 2 \\ &\log 2 - \frac{e^{2K\hat{y}_u}}{2K} = 0 \Rightarrow \hat{y}_u = \frac{\log\left(2K\log 2\right)}{2K} \end{aligned} \tag{3.33}$$

We show in appendix B that these closed-form approximations become more accurate as the sample size increases. In particular, the error of approximating the upper bound by $\hat{y}_u$ approaches zero as $\dfrac{\log\log M}{\log M}$ in the limit $M \to \infty$.

The maximum of the exponent in equation (3.3) is reached at $y^* = \dfrac{\log(q+1)}{2K}$. When $M > M_c(q)$, $y^* \in [y_l, y_u]$ and the estimator coincides with the saddle point approximation of the expected value. Otherwise, the estimate is dominated by y_u.

In summary there is an abrupt change in the behavior of $Z_M^{(q)}$,

$$Z_M^{(q)} \doteq \begin{cases} \exp\left[\frac{(q+1)}{2}\left(\log(q+1) - 1\right)\right], & \lambda(M,q) \geq 1 \\ \exp\left[Kqy_u - K\log 2\right], & \lambda(M,q) \leq 1 \end{cases} \underset{y_u \approx \widehat{y}_u}{\approx} \tag{3.34}$$

$$\approx \begin{cases} \exp\left[\frac{(q+1)}{2}\left(\log(q+1) - 1\right)\right], & \widehat{\lambda}(M,q) \geq 1 \\ \exp\left[\frac{q}{2}\log\left(2\log M\right) - \log M\right], & \widehat{\lambda}(M,q) \leq 1 \end{cases} \tag{3.35}$$

where $\lambda(M,q)$ needs to be computed by numerical methods. A closed-form expression using the approximation $\hat{y}_u$ is $\widehat{\lambda}(M,q) = \dfrac{2\log M}{q+1}$.

An approximate closed-form expression for the maximum bias $B_{max}^{(q)} = \log \mathbb{E}\left[X^q\right] - q\mathbb{E}\left[\log X\right]$ can also be given by computing

$$\mathbb{E}\left[\log X\right] = \frac{1}{2}(-\gamma - \log 2) \tag{3.36}$$

$$\mathbb{E}\left[X^q\right] = \frac{2^{q/2}\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} \tag{3.37}$$

$$B_{max}^{(q)} = \frac{q\gamma}{2} + \log\left(\frac{\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}\right) \tag{3.38}$$

where γ is the Euler-Mascheroni constant and Γ is the gamma function.

Some values of M_c as a function of the order of the moment q are shown in table 3.2. It can be observed that critical sizes are much smaller than in lognormal case.

Table 3.2: Critical sample size M_c for some values of q

Shape parameter σ	Critical size (M_c)
2	1.21
4	12, 18
8	90.01
16	4914.76

Finally, the validity of the analysis carried out is illustrated by comparing the asymptotic results and simulations in terms of $B_M^{(q)}/B_{max}^{(q)}$ (figure 3.3). It can be observed that the closed-form approximation derived in the perturbative analysis models the estimator behaviour more accurately than the numerical approximation.

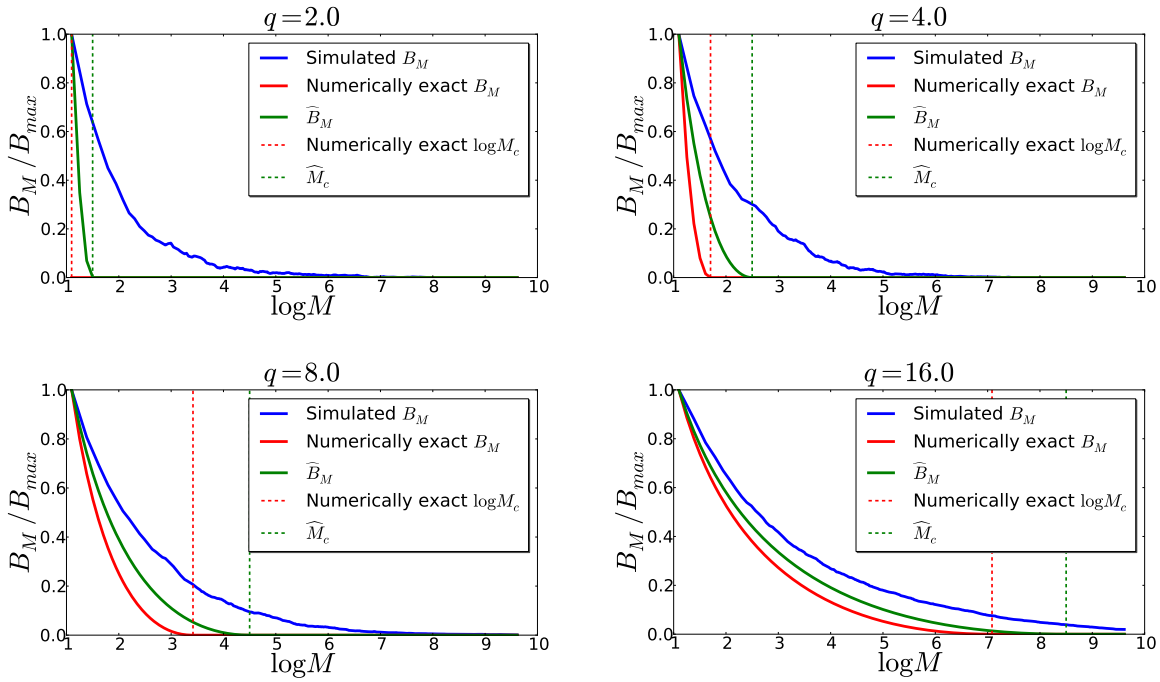


Figure 3.3: Comparison of the described models for Normal case to simulations.

For low orders, $q = 2, 4$, both models produce large errors whereas for higher orders the accuracy increases significantly. In contrast to the lognormal and Log-Exponential-Power-Law distributions, the transition point is reached even for vales of q as large as $q = 16$. One can also concludes from the results of the simulations that the results obtained from the balancing of the dominant terms in the solution of $s_a(y) = 0$ yields a better approximation than the exact numerical solution of this equation.

3.4 The Weibull distribution

In this section we consider the Weibull distribution, also known as the stretched exponential distribution [20]. The Weibull distribution has been used in different contexts: for example, to model the distribution of wind speeds [21] or in survival analysis [22]. It is also the distribution of powers of exponential random variables.

Let Z be an exponential random variable of parameter $\lambda = 1$. The probability density function is

$$f_Z(z) = e^{-z}, \quad z \in [0, \infty). \tag{3.39}$$

Define the random variable $X = Z^q$. The corresponding density is

$$f_X(x) = \frac{1}{q} x^{\frac{1}{q}-1} \exp\left[-x^{1/q}\right], x \in [0, \infty) \tag{3.40}$$

Therefore, X is a Weibull random variable with parameter $1/q$. The other usual parameter in the Weibull distribution is not present because, without loss of generality, the scale parameter in the exponential distribution has been fixed to $\lambda = 1$. The moments of exponential random variables are simply the mean of a Weibull random variable with shape parameter $1/q$. Consider now $Y = \log X$. For this new random variable

$$f_Y(y) = \frac{1}{q} \exp\left[-e^{y/q} + \frac{y}{q}\right], y \in (-\infty, \infty) \tag{3.41}$$

The corresponding entropy function,

$$s_a(y) = \log 2 - \frac{e^{Ky/q}}{K} + \frac{y}{q}, \tag{3.42}$$

with $K = \log 2 / \log M$. Since the entropy function is non-negative, it is only defined for $y \in [y_l, y_u]$. However, $s_a(y) = 0$ is an implicit equation for these bounds. Therefore closed formulas for y_l and y_u cannot be obtained. A solution is the numerical approximation of the boundaries. Another possibility is to make a perturbative analysis based on balancing the leading terms of the equation as $M \to \infty$

$$\log 2 + \frac{\hat{y}_l}{q} = 0 \Rightarrow \hat{y}_l = -q \log 2 \tag{3.43}$$

$$\log - \frac{e^{K\hat{y}_u/q}}{K} = 0 \Rightarrow \hat{y}_u = \frac{q \log (K \log 2)}{K} \tag{3.44}$$

The error in $\hat{y}_u$ tends to zero as $\dfrac{\log \log M}{\log M}$ (see appendix C) as in the folded normal case. By proposition 2.2.1, we can distinguish two regimes, depending on the location of the maximum of the exponent $y^* = \frac{q \log(q+1)}{K}$. When $y^* \in (y_l, y_u)$, Z_M behaves as the asymptotic approximation from saddle point method. If $y_u < y^*$, the estimator is dominated by y_u and finite size sampling effects are important.

The presence of these regimes manifests itself in the sharp change in the behavior of Z_M as a function of sample size

$$Z_M \doteq \begin{cases} \exp\left[-(q+1)\right](q+1)^{q+1}, \ \lambda \geq 1 \\ \exp\left[-\exp\left[\frac{Ky_u}{q}\right] + Ky_u\left(\frac{q+1}{q}\right)\right], \ \lambda \leq 1 \end{cases} \quad \underset{y_u \approx \widehat{y}_u}{\approx} \tag{3.45}$$

$$\approx \begin{cases} \exp\left[-(q+1)\right](q+1)^{q+1}, \ \widehat{\lambda} \geq 1 \\ \exp\left[-\log M + (q+1)\log\log M\right], \ \widehat{\lambda} \leq 1 \end{cases} \tag{3.46}$$

where $\hat{\lambda}(M,q) = \dfrac{\log M}{q+1}$ is an approximation of the exact function $\lambda(M,q)$. Hence, a second order phase transition in $B_M^{(q)}$ appears in the limit $M \to \infty$, $q \to \infty$ and $\frac{\log M}{q} \to$ constant.

The maximum value of the bias $B_{max}^{(q)}$ can be explicitly computed,

$$\mathbb{E}\left[\log X\right] = -q\gamma \tag{3.47}$$

$$\mathbb{E}\left[X\right] = q! \tag{3.48}$$

$$B_{max}^{(q)} = \log q! + q\gamma \tag{3.49}$$

Some values of $M_c(q)$ as a function of the order of the moment q are given in table 3.3.

Table 3.3: Critical sample size M_c for some values of q

Shape parameter σ	Critical size (M_c)
2	20.08
4	148.41
8	8103.08
16	2.41×10^7

Finally, the validity of this analysis is illustrated in simulation experiments. The dependence of the bias of $\log Z_M$ as a function of the size of the sample is displayed in figure 3.4 for different values of q.

It can be observed that both models produce, in general, smaller errors than in Folded-Normal case (figure 3.3). As in previous we obtain the perturbative analysis provides a more accurate approximation. Eventually, note that transition point is not reached for $q = 16$.

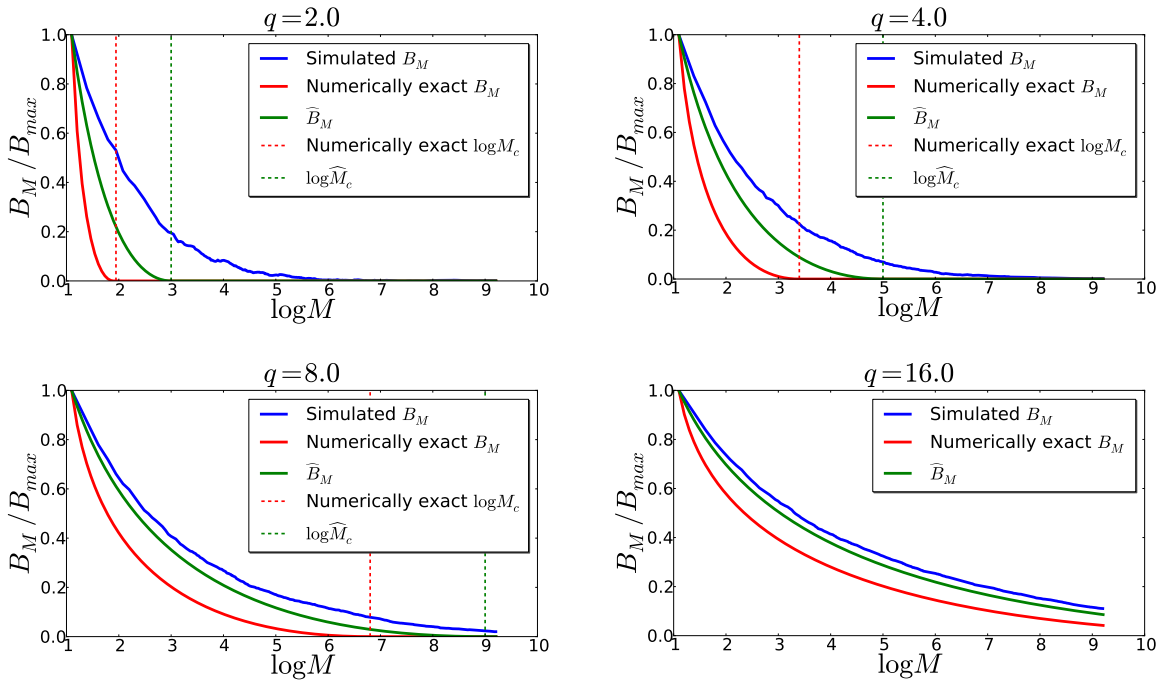


Figure 3.4: Comparison of the described models for the Weibull distribution and the results of simulations.

3.5 The Chi-Squared distribution

As in previous cases, we are going to carry out a similar analysis for the Chi-Squared distribution. The Chi-Squared distribution is commonly used is statistical tests in goodness-of-fit and independence tests [23]. It is also the distribution of sums of squared Normal random variables,

$$X = \sum_{i=1}^{d} Z_i^2, \quad Z_i \sim \mathcal{N}(0,1), \quad X \sim \chi(d). \tag{3.50}$$

The pdf of a Chi-squared random variable with d degrees of freedom ($X \sim \chi(d)$) is

$$f_X(x) = \frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} \exp\left[-x/2\right], x \in [0,\infty), d \in \mathbb{N}^+ \tag{3.51}$$

Consider now the random variable $Y = \log X$, whose probability density function is

$$f_Y(y) = C_d \exp\left[-\frac{e^y}{2} + \frac{d}{2}y\right], y \in (-\infty,\infty) \tag{3.52}$$

In this case, the entropy function is

$$s_a(y) = \log 2 - \frac{e^{Ky}}{2K} + \frac{d}{2}y, \tag{3.53}$$

in the region in which this expression is non-negative. The value of the entropy is zero everywhere except in the interval $[y_l, y_u]$, where y_l and y_u are defined as the solutions of the nonlinear equation

$$s_a(y_l) = 0, \quad s_a(y_u) = 0, \quad y_l < y_u. \tag{3.54}$$

As in the cases of a folded Gaussian and of the Weibull distribution, it is not possible to give exact expressions for the bounds of the interval in which the entropy is non-zero. Nonetheless, the values of y_l and y_u can be estimated using numerical algorithms. Alternatively, closed-form approximate expressions can be obtained by balancing the two leading terms in the nonlinear equation

$$\log 2 + \frac{d}{2}\hat{y}_l = 0 \Rightarrow \hat{y}_l = -\frac{2}{d}\log 2 \tag{3.55}$$

$$\log 2 - \frac{e^{K\hat{y}_u}}{2K} = 0 \Rightarrow \hat{y}_u = \frac{\log(2K\log 2)}{K} \tag{3.56}$$

The difference between $\hat{y}_u$ and y_u approaches zero as $\dfrac{\log\log M}{\log M}$ when $M \to \infty$.

An abrupt change of behavior in $Z_M^{(q)}$ can be found depending on the location of the local maximum of the exponent in proposition 2.2.1. For small samples ($M < M_c(d,q)$) the local maximum is beyond the interval $[y_l, y_u]$ and the maximum in the relevant interval $[y_l, y_u]$ is $y^* = y_u$. In this regime, effects related to the size of the sample are dominant in the average. For large samples, if $M > M_c(d,q)$, the dominant contribution comes from the maximum y^* and the estimator is well approximated by the result of saddle point integration.

By computing the maximum in (2.2.1) the empirical moment estimator is to exponential leading order

$$Z_M^{(q)} \doteq \begin{cases} \exp\left[-\left(\frac{d}{2}+q\right) + \left(\frac{d}{2}+q\right)\log\left(d+2q\right)\right], & \lambda(M,d,q) \geq 1 \\ \exp\left[\frac{e^{Ky_u}}{2} + K\left(\frac{d}{2}+q\right)y_u\right], & \lambda(M,d,q) \leq 1 \end{cases} \underset{y_u \approx \widehat{y}_u}{\approx} \tag{3.57}$$

$$\approx \begin{cases} \exp\left[-\left(\frac{d}{2}+q\right) + \left(\frac{d}{2}+q\right)\log\left(d+2q\right)\right], & \widehat{\lambda}(M,d,q) \geq 1 \\ \exp\left[-\log M + \log\left(2\log M\right)\left(\frac{d}{2}+q\right)\right], & \widehat{\lambda}(M,d,q) \leq 1, \end{cases} \tag{3.58}$$

where $\widehat{\lambda}(M,d,q) = \dfrac{2\log M}{d+2q}$ is the function that determines where the transition occurs.

Note that using the approximate value $\hat{y}_u$, one recovers the results for the normal distribution with $q = 2$ using the results for the Chi-Squared distribution with $d = 1, q = 1$

$$\widehat{\lambda}_{\text{Chi-squared}}(d=1,q=1) = \widehat{\lambda}_{\text{Normal}}(q=2) = \frac{2\log M}{3}. \tag{3.59}$$

As we did in previous cases, the following table contains some values of M_c as a function of the order of the moment q.

Table 3.4: Critical sample size M_c for some values of q with 5 degrees of freedom

Shape parameter σ	Critical size (M_c)
2	90.01
4	665.14
8	36315.5
16	1.08×10^8

An explicit expression can be obtained for the maximum bias $B_{max}^{(q)} = \log \mathbb{E}\left[X^q\right] - q\mathbb{E}\left[\log X\right]$,

$$\mathbb{E}[X^q] = 2^q \frac{\Gamma\left(q + \frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{3.60}$$

$$\mathbb{E}[\log X] = \psi\left(\frac{d}{2}\right) + \log 2 \tag{3.61}$$

$$B_{max}^{(q)} = \log\left(\frac{\Gamma\left(q + \frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right) - q\psi\left(\frac{d}{2}\right) \tag{3.62}$$

where $\psi(x)$ and $\Gamma(x)$ are the gamma and the digamma functions respectively.

The validity of this analysis is illustrated using simulations. The dependence of the normalized bias of $\log Z_M^{(q)}$ as a function of M, the sample size, is depicted in figure 3.5. The most striking feature of these plots is the extremely accurate approximation that results from using $\hat{y}_u$ instead of the exact y_u.

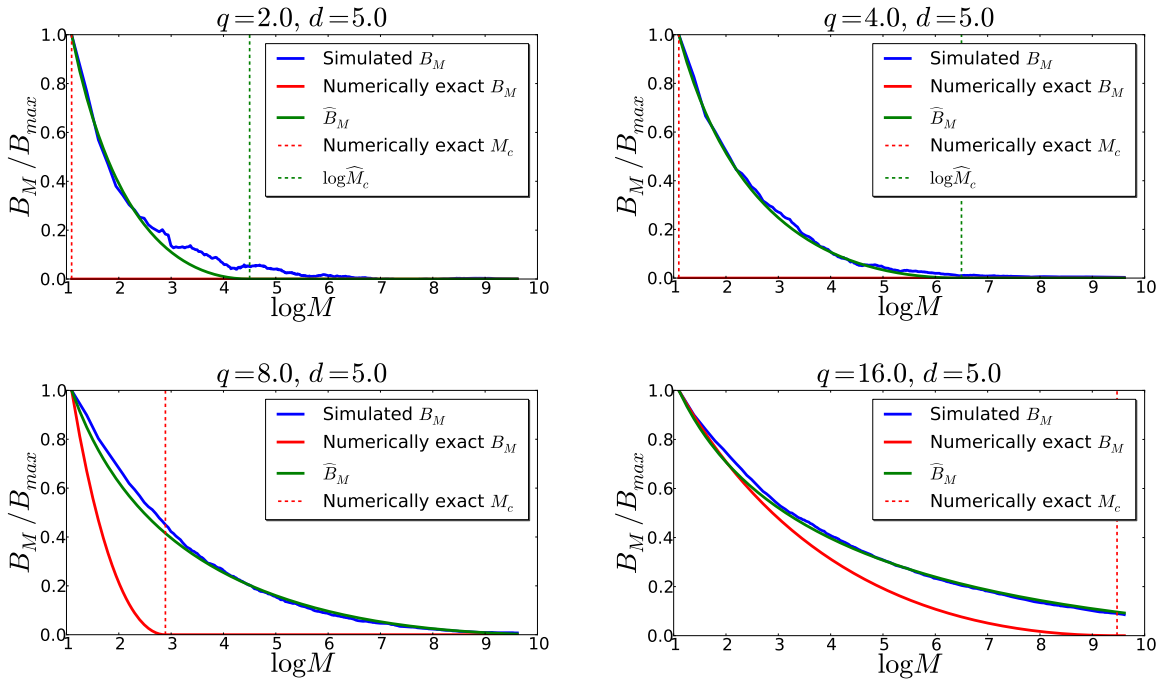


Figure 3.5: Comparison of the described models for Chi-Squared distribution to simulations, $d = 5$.

3.6 Critical sample size

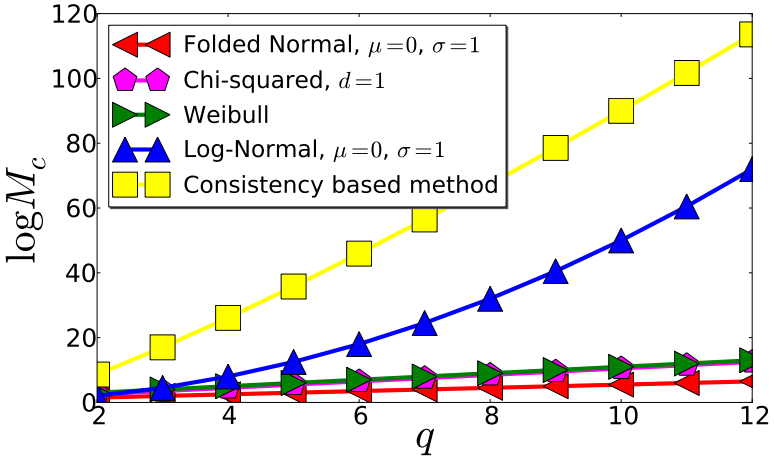


Figure 3.6: Critical sizes for analyzed cases.

In this section we compare the critical sizes for the moment estimators for different distributions that have been computed in previous sections. The results of this comparison are compiled in Table 3.5 and depicted in Figure 3.6. From this figure, it is apparent that the

Table 3.5: Lambda functions for the analyzed distributions

Distribution	$\lambda(M,q)$	$M_c(q=2)$	$M_c(q=4)$
Folded Normal $\mu=0,\sigma=1$	$\frac{2\log M}{q+1}$	4.48	12.18
Chi-Squared (d)	$\frac{2\log M}{2q+d}$	12.18	90.01
Weibull	$\frac{\log M}{q+1}$	20.08	148.41
lognormal, $\mu=0,\sigma=1$	$\frac{2\log M}{q^2}$	7.39	2980.95
Log-Exponential-Power-Law, $\rho=2$	$\sim \frac{\log M}{q^{\rho/(\rho-1)}} = \frac{\log M}{q^2}$	54.59	8.88×10^6
Consistency-based method	$\frac{\log M}{2q\log\log M}$	5503.66	2.14×10^{11}

value $M_c(q)$ strongly depends on the probability distribution considered. In particular, the lognormal distribution has higher critical sizes than the other distributions. This is consistent with the fact that the tails of the lognormal distribution are heavier. The estimates of the critical sample size based on requiring consistency (see section 2.1) are exceedingly high.

It is interesting to note that, for large q and M, $\lambda(M,q) = \text{constant}\frac{\log M}{q^n}$ for different values of $n > 0$.

The results of the empirical analysis presented in this section illustrate that the approximations obtained for the critical size $M_c(q)$ work well for sufficiently heavy-tailed different distributions. Poor estimations are obtained for low-order moments of random variables with exponential decay, as illustrated by the results of simulations presented in figure 3.7.

Our hypothesis is the developed analysis are only applicable for heavy-tailed distributions that do not have algebraic decay. Note that even if the distribution of the original variable is not heavy-tailed, powers of the original variable will eventually be heavy-tailed. These observations imply that the change in regime observed in the sample estimator appears in very general situations.

Note also that the lambda function of the Log-Exponential-Power-Law distributions

$$\lambda(M,q,\rho) \sim \frac{\log M}{q^{\rho/(\rho-1)}} \tag{3.63}$$

interpolates between different cases:

1. As we mentioned in section 3.2, the lognormal case is recovered for $\rho = 2$.
2. If $\rho \longrightarrow \infty$,

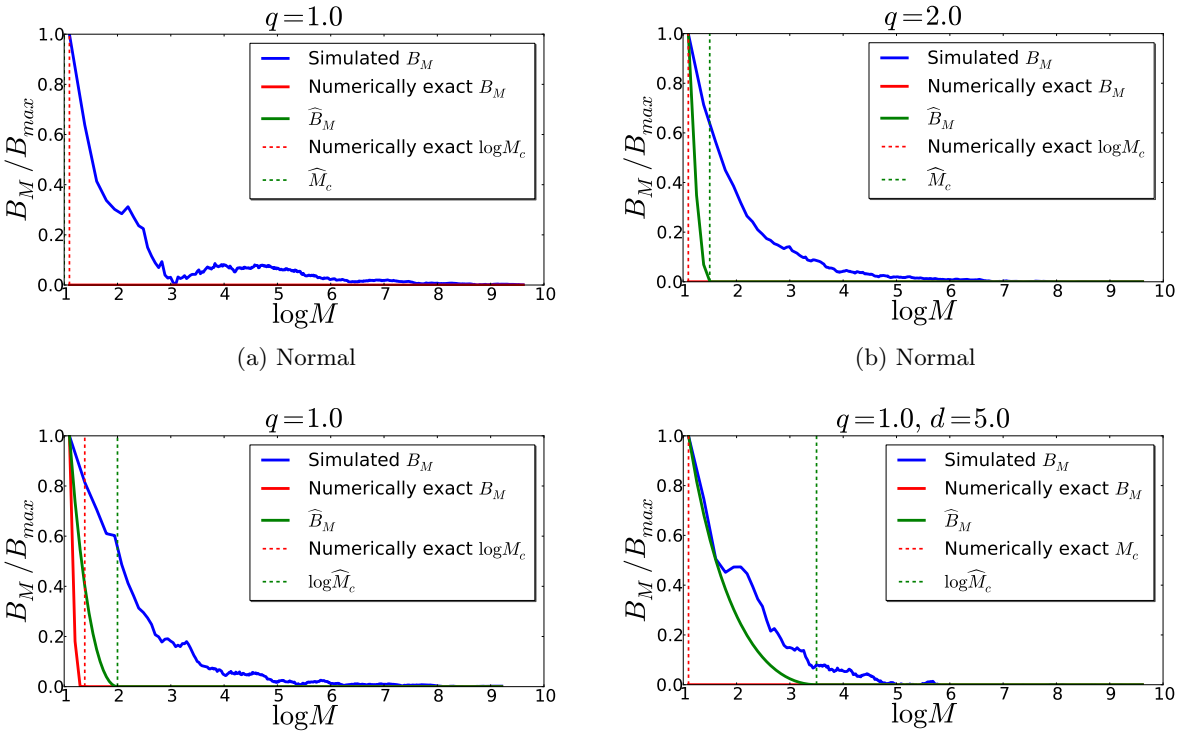
$$\lambda(M,q,\rho) \sim \text{constant}\ \ \frac{\log M}{q} \tag{3.64}$$

which is, neglecting the constants, the same λ function as for the Folded-Normal, Chi-Squared and Weibull cases.

3. If $\rho \longrightarrow 1^{+}$,

$$\log M_c(q,\rho) \sim q^{\rho/(\rho-1)} \longrightarrow \infty \tag{3.65}$$

that is, we cannot reach the asymptotic regime since $M_c(q,\rho)$ increases to infinity. This situation is found in the analysis of Pareto distributions (see next section).



(a) Normal

(b) Normal

(c) Weibull

(d) Chi-Squared

Figure 3.7: Low order moments modelization.

3.7 The Pareto distribution

The Pareto distribution appears in many different contexts. It is the distribution of degrees of nodes in scale-free networks [24], the distribution of the frequency of words in texts [25], the time between large earthquakes [26], etc. In all these cases, averages of random samples are needed and it is reasonable to try a similar approach as in the previous cases.

Let X be a random variable that follows a Pareto distribution

$$f_X(x) = \alpha x^{-\alpha+1}, x \in [1, \infty) \tag{3.66}$$

The pdf of the random variable $Y = \log X$ is

$$f_Y(y) = \alpha \exp\left[-\alpha y\right], y \in [0, \infty) \tag{3.67}$$

Solving the equation (2.22) for this pdf,

$$s_a(y) = \log 2 - \alpha y \tag{3.68}$$

From the functional form of the entropy function it is possible to see that the maximum of the exponent in (2.26) is $y_l = 0$, the lower bound of the interval, in which the entropy is non-zero. Therefore, the entropy (rate in LDT) function is zero to leading exponential order, and

$$Z_M \doteq \exp\left[0\right]. \tag{3.69}$$

Since a phase transition cannot be obtained, the asymptotic regime is never reached.

Based on the ideas in [19], we can use a truncated moments approach,

$$Z_M \approx \int_0^{U_M} x \; f_X(x) dx \tag{3.70}$$

where U_M, the truncation value, depends on M, the sample size.

A first approach is to use a statistic of G_M, the maximum of the sample of size M. The probability distribution and corresponding density are

$$\begin{aligned} F_{G_M}(u) &= \mathbb{P}\left(G_M < u\right) = \left(\mathbb{P}(X < u)\right)^M = F_X(u)^M \\ f_{G_M}(u) &= M F_X(u)^{M-1} f_X(u). \end{aligned}$$

For a Pareto distribution, the quadrature (3.70) can be performed explicitly

$$Z_M \approx \int_1^{U_M} \alpha x^{-\alpha} dx = \frac{\alpha}{\alpha - 1}\left(1 - U_M^{-\alpha+1}\right) \tag{3.71}$$

We are going to note as $U_M^{(i)}, i = 1, 2, 3$ each choice of the truncation value. A reasonable truncation value could be $U_M^{(1)}$ as the mode of the distribution of the maximum. For a continuous distribution, the mode is the maximum of the pdf,

$$f'_{G_M}(U_M^{(1)}) = 0 \Leftrightarrow (M-1) f_X(U_M^{(1)})^2 + F_X(U_M^{(1)}) f'_X(U_M^{(1)}) = 0 \tag{3.72}$$

In case of the Pareto distribution,

$$(M-1)\alpha {U_M^{(1)}}^{-\alpha} - (1-x^{-\alpha})(\alpha+1) = 0 \;\Rightarrow\; U_M^{(1)} = \left(\frac{\alpha+1}{M\alpha+1}\right)^{-1/\alpha} \tag{3.73}$$

For other types of distributions, this value may need to be estimated numerically. Alternatively, to obtain closed-form expressions, one can approximate the distribution of the maximum by using Extreme Value Theory (EVT) [14]. The Fisher-Tippett-Gnedenko [27, 28] theorem states that the distribution of the maximum of any distribution behaves asymptotically as either a Gumbel distribution, a Weibull distribution or a Fréchet distribution. This theorem can be expressed in a compact way using $H_\xi(x)$, the Generalized Extreme Value Distribution (GEV).

Theorem 3.7.1 *Let $\{X_i\}$ be a sequence of i.i.d random variables and $G_M = \max_i \{X_i\}$. There exist norming constants b_M and c_M such as*

$$\lim_{M\to\infty} \mathbb{P}\left(\frac{G_M - b_M}{c_M} < x\right) = H_\xi(x) = \begin{cases} \exp\left[-(1+\xi x)^{-1/\xi}\right] \;,\; \xi \neq 0, \\ \exp\left[-e^{-x}\right] \;,\; \xi = 0 \end{cases} \tag{3.74}$$

Choosing the parameter ξ, $H_\xi(x)$ takes the form of one of the three distributions mentioned before. Applying

$$F_{G_M - b_M/c_M}(u) = F_{G_M}(C_M u + b_M) \tag{3.75}$$

we will have a closed-form for the asymptotic distribution of the maximum.

Depending on the distribution of X, the distribution of the sample maximum approaches one of these three distributions as the size of the sample increases. In every case, the mode and the mean have closed formulas thus, even in complicated distributions, we can have closed-form approximations. In particular, for Pareto random variables, the distribution of the maximum belongs to the domain of attraction of the Fréchet distribution

$$c_M^{-1} G_M \xrightarrow{d} \Phi_\alpha \tag{3.76}$$

with $c_M = M^{1/\alpha}$ and $\Phi_\alpha(x) = \exp\left[-x^{-\alpha}\right]$ the cdf of Fréchet distribution.

Computing the probability for the maximum of Pareto random samples,

$$\mathbb{P}(G_M < u) = \exp\left[-\frac{x^{-\alpha}}{M}\right] \tag{3.77}$$

One can use the mean or the mode of the Fréchet distribution as truncation values in (3.71)

$$U_M^{(2)} = \mathbb{E}[G_M] = M\Gamma\left(1 - \frac{1}{\alpha}\right) \tag{3.78}$$

$$U_M^{(3)} = \text{mode}[G_M] = M\left(\frac{\alpha}{1+\alpha}\right)^{1/\alpha} \tag{3.79}$$

Evaluating in every of the three given solutions,

(3.80)

$$U_M^{(1)} \text{ (Mode of the distribution of the maximum)} \Rightarrow Z_M \approx \frac{\alpha}{\alpha-1}\left(1-\left(\frac{\alpha+1}{M\alpha+1}\right)^{\frac{\alpha-1}{\alpha}}\right)$$

(3.81)

$$U_M^{(2)} \text{ (Mean of the EVT limit distribution)} \Rightarrow Z_M \approx \frac{\alpha}{\alpha-1}\left(1-\left[M\Gamma\left(1-\frac{1}{\alpha}\right)\right]^{1-\alpha}\right)$$

(3.82)

$$U_M^{(3)} \text{ (Mode of the EVT limit distribution)} \Rightarrow Z_M \approx \frac{\alpha}{\alpha-1}\left(1-M^{1-\alpha}\left(\frac{\alpha}{\alpha+1}\right)^{1-\alpha/\alpha}\right)$$

We test the validity of the three different choices for Pareto distribution in figure 3.8. It can be observed that the approximation by choosing U_M as the mode of the explicit distribution of the maximum is the most accurate of the three options. This figure illustrates that considering the explicit distribution of the maximum is better than the asymptotic one. However, this approach might not be used in all the cases and EVT could be applied to model the estimator behaviour.

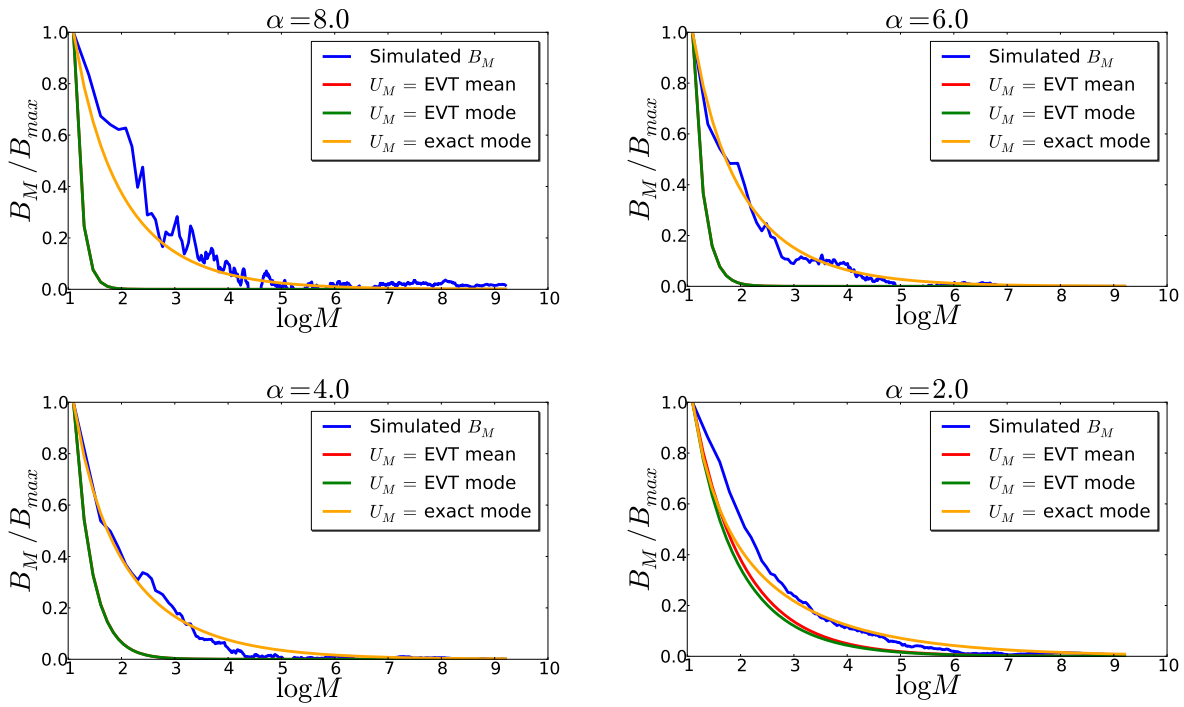


Figure 3.8: Bias modelization for each of the three approaches for Pareto distribution. EVT mean and EVT mode models match almost perfectly for the graphics with $\alpha > 4$.

Chapter 4

Conclusions and further work

Conclusions

In this work we have analyzed the convergence of sample averages to the actual value of mean, which, according to the law of large numbers, is the asymptotic limit of this quantity as the sample size increases. The estimator is unbiased, because the expected value of the sample average coincides with the mean for any sample size. However, the distribution of this estimator for small samples can in general be very skewed. Specifically, its maximum, which corresponds to *typical* values of the estimator, can be quite different from its mean. This difference between typical and average behavior means that, in practice, the estimator, in spite of being unbiased, is not accurate for small samples. As the sample size increases the typical values of the estimator approach the expected vale. Asymptotically, for distributions of finite variance, the estimator has a normal distribution, centered around the actual value of the mean and a variance that is inversely proportional to the sample size. To describe how this transition takes place, we have performed an analysis in terms of the bias of the logarithm of the empirical estimator. As shown in the introduction, this quantity has some desirable properties such as monotonicity and asymptotic convergence (see propositions (1.1.3) and (1.1.1) respectively). These properties make the bias of the logarithm of the estimator a useful tool to analyze the behavior of the estimator as a function of the sample size.

The analysis of the logarithm bias of the estimator is inspired by the analysis of the partition function in models of disordered systems in the area of statistical physics. Specifically, we have made extensive use of the analysis made in of the random energy model [10]. The partition function in the random energy model is a sum of exponentials of minus β, the inverse temperature, times a normal random variable whose variance scales with the number of energy levels (the sample size in our problem). The *quenched* free energy is proportional to the expectation of the logarithm of the system's partition function Under these conditions one finds a phase transition as a function of temperature β^{-1} (2.29). For low temperatures $\beta > \beta_c$, the sample free energy is different from its asymptotic value. As the temperature of the system increases, in the limit in which the number of energy levels approaches infinity (limit of large samples), the bias in the quenched free energy decreases, and, above a transition temperature β_c^{-1}, becomes exactly zero. The transition is marked by a discontinuity in the second derivative of the quenched free energy. The analysis made in REM is not completely formal. Nonetheless, it is possible to provide a more rigorous derivations of REM's results within the mathematical

framework of Large Deviation Theory.

In chapter 3 the tools employed in the REM are applied to the analysis of the convergence of averages of independent identically distributed random variables sampled from different distributions. The main difference with the analysis performed in REM is that the parameters of this distributions do not depend on the size of the sample. For this reason, we focus on the analysis of sample estimates of the qth moment. The main result of this analysis is that, for a wide class of distributions, the logarithm of the empirical moment exhibits a phase transition akin to the one found in REM in the limit of large samples $M \to \infty$ and large $q \to \infty$ with $\log M/q^n$ finite for some n that depends on the particular distribution considered. The analysis is made for sample estimates of the moments of lognormal, exponential, normal, Chi-Squared and Weibull random variables. In all the studied cases, the analysis is validated using Monte-Carlo simulations.

As discussed in section 2.2.1, the REM and LDT analysis can only be applied to certain types of distributions. It cannot be applied to distributions whose tails decay algebraically, such as the Pareto distribution. For these types of distributions, the entropy function is zero. Therefore, for this case, the convergence of sample moment estimator is analyzed in terms of estimations of truncated moments. The truncation threshold is chosen as a statistic (the mode or the mean) of the distribution of the sample maximum 3.7. In contrast to the previous case, there is no phase transition: one is always in the small sample regime. Monte Carlo simulations are used to illustrate the results of the analysis performed.

Further work

As shown in section 2.2.1, Large Deviation Theory provides a formal framework for the random energy model analysis. One of the lines of future research is to provide the theoretical justification of the results shown in chapter 3 using the LDT framework. In particular, one can apply the Large Deviation Principle to the measure (2.55) associated with different distributions and then use Varadhan's Theorem to obtain corresponding rate function. The truncated moments estimate developed for the Pareto distribution in section 3.7 could also prove a useful tool for analysis of the other distributions.

In the random energy model there is a second phase transition for the fluctuations around the sample average [29]. Presumably, this second transition is also present for the distributions studied in this work. One may expect the existence of three separate regimes: the first one (small samples), in which where the logarithm of the sample average is a biased estimator of the logarithm of the mean; a second regime in which the logarithm of the sample average is unbiased but where the logarithm of the sample variance is a biased estimator of the logarithm of the variance; finally, for large samples, a regime in which these two quantities have converged and the CLT applies, so that the empirical average is approximately normal.

From the observations made in section 3.6, we hypothesize that the REM analysis is valid only for subexponential distributions [30, 31] with non-algebraic tail decay. A random variable

X is said to be subexponential if

$$\frac{\mathbb{P}\left(X_1 + ... + X_M < x\right)}{M\mathbb{P}\left(X < x\right)} \underset{x\to\infty}{\longrightarrow} 1 \tag{4.1}$$

It is worth investigating whether this is the relevant asymptotic regime for the analysis of finite sample averages.

The quantity $\log Z_M^{(q)}$ is strongly related to cumulant generating function. To make this connection apparent, define the random variable $Y = \log X$

$$\log Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} \exp\left[qY\right] \underset{M\to\infty}{\longrightarrow} \log \mathbb{E}[e^{qY}] = g(q) \tag{4.2}$$

where the $g(q)$ is

$$g(q) = \sum_{i=1}^{\infty} \kappa_i \, \frac{q^i}{i!} \tag{4.3}$$

and κ_i is the cumulant of order i.

Another possible approach to analyze the behavior of the sample moment estimator is to study the convergence properties of the Edgeworth Series [32] of a given distribution. Let Z_M be the sum of standardized M random variables. The Central Limit Theorem states that

$$\lim_{M\to\infty} F_{Z_M}(x) = \Phi(x) \tag{4.4}$$

where $\Phi(x)$ is the cdf of the Normal distribution. From this relation one obtains

$$F_{Z_M}(x) = \Phi(x) + \sum_{j=1}^{\infty} \frac{P_j(-\partial_x)}{M^{j/2}}\Phi(x), \tag{4.5}$$

wher $\partial_x \equiv \frac{\partial}{\partial x}$ is the derivative operator The summation in this formula indicates how far the estimator is from the Normal distribution and the coefficients of the polynomials $P_j(\cdot)$ can be written in terms of the cumulants. A line of research is the analysis of the convergence of the sample estimates of cumulants and, as a result of this analysis, establish the conditions under which the CLT is an accurate approximation for Z_M.

The results for lognormal random variables can be applied to the field of telecommunications [33]. In particular, lognormal distribution has been used as a model in slow fading in communications channels. The performance of wireless systems performance with interfering stations can be studied in terms of sums of lognormal random variables. The goal of most previous studies in this area [33, 34] is to approximate the distribution of these sums.

In statistics and machine learning sums of iid random variables are ubiquitous. For instance, maximum likelihood estimators in model fitting are generally expressed in terms of sample averages

$$\widehat{\Theta}_{ML} = \sum_{i=1}^{M} g(x_i) \tag{4.6}$$

Analyzing the distribution of $\{g(x_i)\}$ it is possible to determine a critical size M_c such that, for sufficiently large samples $M > M_c$, the maximum likelihood estimator (4.6) is accurate. For instance, a maximum likelihood fit of the tail index of a Pareto distribution, α, from a sample $\{x_i\}_{i=1}^{M}$, yields

$$\widehat{\alpha} = 1 + M \left[\sum_{i=1}^{M} \log x_i \right]^{-1} . \tag{4.7}$$

$\widehat{\alpha}$ is the so called Hill estimator [35].

In summary, the main contribution of this work is to propose a criterion for the convergence of typical values of the sample average to the expected value in terms of the bias of the logarithm of this estimator. Since sums of random variables appear in a vast number of scientific fields, these results have a wide range of application. In particular, they open new lines of research in computational statistics and in machine learning.

Appendix A

Lognormal distribution

The pdf of a lognormal random variable X is a lognormal random variable if and only if its pdf is given by

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], x \in (0, \infty) \tag{A.1}$$

X can be expressed as the exponential of a normal random variable,

$$X = e^{\mu+\sigma Y}, Y \sim \mathcal{N}(0, 1) \tag{A.2}$$

The parameter e^{μ} is a scale parameter. Therefore one can set $\mu = 0$ without loss of generality. All the moments of the lognormal distribution are defined.

If X is lognormal with parameters $\mu = 0$ and σ, X^q also follows a lognormal distribution with parameters $\mu = 0$ and σq. In consequence, one only needs to carry out the analysis for a lognormal distribution with parameters $\mu = 0$ and σ.

A direct and asymptotically exact estimate using the saddle point approximation can be applied

$$\mathbb{E}[e^Y] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2} + y\right] dy \doteq \exp\left[\max_{(-\infty,\infty)}\left\{-\frac{y^2}{2\sigma^2} + y\right\}\right] \doteq \exp\left[\frac{\sigma^2}{2}\right] \tag{A.3}$$

Let's assume we have a random sample of size $M = 2^K$ distributed as X. We can express the probability of having a value of the sample in an interval $[K\epsilon, K(\epsilon + \delta)]$ in terms of the entropy function $s_a(y)$

$$\mathbb{P}\left(Y \in [K\epsilon, K(\epsilon+\delta)]\right) = \int_{K\epsilon}^{K(\epsilon+\delta)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-y^2}{2\sigma^2}\right] dy = \int_{\epsilon}^{\epsilon+\delta} \frac{K}{\sqrt{2\pi\sigma^2}} \exp\left[K\left(\frac{-Kz^2}{2\sigma^2}\right)\right] dz \doteq$$
$$\doteq \exp\left[K\left(\max_{y\in(\epsilon,\epsilon+\delta)}\left\{-\frac{Ky^2}{2\sigma^2}\right\}\right)\right] \doteq \exp\left[K\left(\max_{y\in(\epsilon,\epsilon+\delta)}\{s_a(y) - \log 2\}\right)\right]$$

where equations (2.18) and (2.20) have been combined. The entropy function is then given by

$$s_a(y) = \log 2 - \frac{Ky^2}{2\sigma^2} \tag{A.4}$$

$s_a(y)$ is a non-negative function, so it is only defined for an interval of possible values of y, $|y| \leq \sqrt{\frac{2\sigma^2 \log 2}{K}}$. We will note as y_l and y_u the lower bound and the upper one respectively.

$$Z_M \approx \frac{1}{M} \int_{y_l}^{y_u} \exp\left[K\left\{s_a(y) + y\right\}\right] dy \doteq \exp\left[K \max_{[y_l, y_u]} \left\{s_a(y) + y - \log M\right\}\right] \tag{A.5}$$

The maximum of the exponent is reached at $y^* = \frac{\sigma^2}{K}$ and then the estimate is given by evaluating the previous formula at y^*,

$$Z_M \approx \exp\left[\frac{\sigma^2}{2}\right] \tag{A.6}$$

Note this approximation matches the saddle point estimate result. However, this point might not lie in the interval $[y_l, y_u]$. If this occurs, the maximum of the exponent is reached at the upper bound of the interval y_u. Evaluating,

$$Z_M \approx \exp\left[\sqrt{2K\sigma^2 \log 2} - K \log 2\right] = \exp\left[\sqrt{2\sigma^2 \log M} - \log M\right] \tag{A.7}$$

Analyzing lognormal distribution we have found two cases with different behavior. First one corresponds to $M \geq M_c(\sigma)$, where the best estimate is given by the saddle point method. The second case, $M_c(\sigma) \leq M$ depends on the sample size and is biased compared to saddle point estimate. The critical point where *phase transition* occurs, $y^* = y_u$,

$$K = \frac{\sigma^2}{2 \log 2} \Leftrightarrow M = \exp\left[\frac{\sigma^2}{2}\right] \tag{A.8}$$

From this result we can describe the phase transition in terms of the sample size

$$Z_M \doteq \begin{cases} \exp\left[\frac{\sigma^2}{2}\right], \lambda(M, \sigma) > 1 \\ \\ \exp\left[\sqrt{2\sigma^2 \log M} - \log M\right], \lambda(M, \sigma) < 1 \end{cases}$$

where $\lambda(M, \sigma) = \frac{2 \log M}{\sigma^2}$

Appendix B

Folded Normal distribution

The pdf of a folded normal (or folded Gaussian) random variable X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] + \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(-x-\mu)^2}{2\sigma^2}\right], x \in [0,\infty) \tag{B.1}$$

where μ is a location parameter and σ a scale one. Without loss of generality, one can assume $\mu = 0, \sigma = 1$. The object we want to analyze is the estimator of the qth moment of the distribution

$$Z_M^{(q)} = \frac{1}{M}\sum_{i=1}^{M} X_i^q \tag{B.2}$$

We define the random variable $Y = \log X$ whose pdf is

$$f_Y(y) = \frac{2}{\sqrt{2\pi}}\exp\left[-\frac{e^{2y}}{2} + y\right] \tag{B.3}$$

A direct saddle point estimation of the q-th moment is

$$\mathbb{E}[e^{qY}] = \int_{-\infty}^{\infty} \frac{2}{\sqrt{2\pi}}\, e^{yq}\, \exp\left[-\frac{e^{2y}}{2} + y\right] dy \doteq \exp\left[\max_{y\in(-\infty,\infty)}\left\{-\frac{e^{2y}}{2} + (q+1)y\right\}\right] =$$

$$= \exp\left[\frac{(q+1)}{2}\left(\,\log(q+1) - 1\,\right)\right] \tag{B.4}$$

Given a interval $[K\epsilon, K(\epsilon+\delta)]$, the probability of it can be estimated by,

$$\mathbb{P}(Y \in [K\epsilon, K(\epsilon+\delta)]) = \int_{K\epsilon}^{K(\epsilon+\delta)} \frac{2}{\sqrt{2\pi}}\exp\left[-\frac{e^{2y}}{2} + y\right] dy = \int_{\epsilon}^{\epsilon+\delta} \frac{2K}{\sqrt{2\pi}}\exp\left[K\left(-\frac{e^{2Kz}}{2K} + z\right)\right] dz \doteq$$

$$\doteq \exp\left[K\left(\max_{y\in[\epsilon,\epsilon+\delta]}\left\{-\frac{e^{2Kz}}{2K} + z\right\}\right)\right] \doteq \exp\left[K\max_{y\in[\epsilon,\epsilon+\delta]}\{s_a(y) - \log 2\}\right]$$

where the term $\log 2$ is a result of the combination of equations (2.18) and (2.20). Solving the equation we get the entropy function is expressed as

$$s_a(y) = \log 2 - \frac{e^{2Ky}}{2K} + y \tag{B.5}$$

We have to find the interval $[y_l, y_u]$ where the entropy is positive. However, this can only be made by numerical methods. Despite this fact, we can make a perturbative analysis by assuming a balance between two dominant terms in the equation when $M = 2^K \to \infty$. For the lower bound of the interval, $\log 2 + \widehat{y}_l = 0$,

$$\widehat{y}_l = -\log 2 \tag{B.6}$$

$$\text{res}_{\widehat{y}_l} = -\frac{e^{-2K\log 2}}{2K} \underset{K\to\infty}{\longrightarrow} 0 \tag{B.7}$$

and the error tends to zero exponentially fast with K.

For the upper bound of the integral, $\log 2 - \dfrac{e^{2K\widehat{y}_u}}{2K} = 0$,

$$\widehat{y}_u = \frac{\log(2K\log 2)}{2K} \tag{B.8}$$

$$\text{res}_{\widehat{y}_u} = \frac{\log(2K\log 2)}{2K} \underset{K\to\infty}{\longrightarrow} 0 \tag{B.9}$$

and, in this case, the error tends much slower to zero, as $\dfrac{\log\log M}{\log M}$.

We can model the behavior of the qth moment estimator by

$$Z_M^{(q)} \approx \int_{y_l}^{y_u} \exp\left[K\left(s_a(y) + qy\right)\right] dy \doteq \exp\left[K \max_{y\in[y_l,y_u]} \{s_a(y) + qy\}\right] \tag{B.10}$$

The exponent has a maximum at $y^* = \dfrac{\log(q+1)}{2K}$ for sufficiently large sample size. In this regime,

$$Z_M^{(q)} = \exp\left[\frac{(q+1)}{2}\left(\ \log(q+1) - 1\ \right)\right] \tag{B.11}$$

as the saddle point estimation of the q-th moment. For small samples the maximum of the exponent in equation (B.10) is reached at the upper bound of the interval,

$$Z_M^{(q)} = \exp\left[Kqy_u - K\log 2\right] \underset{y_u\approx\widehat{y}_u}{\approx} \exp\left[-K\log 2 + \frac{(q+1)}{2}\log(2K\log 2)\right] = \frac{(2\log M)^{(q+1)/2}}{M} \tag{B.12}$$

The critical size that is the border between both regimes is

$$y^* = y_u \Leftrightarrow K_c = \frac{\log(q+1)}{2y_u} \Leftrightarrow M_c = \exp\left[\frac{\log(q+1)}{2y_u}\right] \tag{B.13}$$

$$y^* = \widehat{y}_u \Leftrightarrow \widehat{K}_c = \frac{q+1}{2\log 2} \Leftrightarrow \widehat{M}_c = \exp\left[\frac{q+1}{2}\right] \tag{B.14}$$

by numerical and perturbative analysis respectively. Combining these results, the phase transition between the two regimes,

$$Z_M^{(q)} \doteq \begin{cases} \exp\left[\frac{(q+1)}{2}\left(\log(q+1)-1\right)\right], & \lambda \geq 1 \\ \exp\left[Kqy_u - K\log 2\right], & \lambda \leq 1 \end{cases} \underset{y_u \approx \widehat{y}_u}{\approx} \tag{B.15}$$

$$\approx \begin{cases} \exp\left[\frac{(q+1)}{2}\left(\log(q+1)-1\right)\right], & \widehat{\lambda} \geq 1 \\ \frac{(2\log M)^{(q+1)/2}}{M}, & \widehat{\lambda} \leq 1 \end{cases} \tag{B.16}$$

where $\lambda(M,q)$ can only be computed by numerical methods and $\widehat{\lambda}(M,q) = \dfrac{2\log M}{q+1}$.

Appendix C

Weibull distribution

The pdf of a Weibull random variable X is

$$f_X(x) = \frac{1}{q} x^{\frac{1}{q}-1} \exp\left[-x^{1/q}\right], x \in [0, \infty) \tag{C.1}$$

Without loss of generality, one can set the scale parameter to 1. The shape parameter, usually noted as k, is changed to $1/q$ to emphasize the property of being the distribution of the moments of an exponential random variable with scale parameter 1. So we only have to analyze the mean of this variable to get the behavior of the moments of exponential random variables.

We define the random variable $Y = \log X$ whose pdf is

$$f_Y(y) = \frac{1}{q} \exp\left[-e^{y/q} + \frac{1}{q} y\right], y \in (-\infty, \infty) \tag{C.2}$$

As in the previous cases we can give an estimation of the average by saddle point method,

$$\begin{aligned} \mathbb{E}[e^Y] &= \frac{1}{q} \int_{-\infty}^{\infty} \exp\left[-e^{y/q} + \frac{1}{q} y + y\right] dy \doteq \exp\left[\max_y \left\{-e^{y/q} + \frac{q+1}{q} y\right\}\right] = \\ &= \exp\left[-(q+1)\right] (q+1)^{q+1} \doteq (q+1)! \end{aligned} \tag{C.3}$$

where the last approximation is given by the Stirling's approximation for the factorial.

Computing the probability of the interval $[K\epsilon, K(\epsilon+\delta)]$,

$$\begin{aligned} &\mathbb{P}\left(Y \in [K\epsilon, K(\epsilon+\delta)]\right) = \int_{K\epsilon}^{K(\epsilon+\delta)} \frac{1}{q} \exp\left[-e^{y/q} + \frac{1}{q} y\right] dy = \\ &= \frac{K}{q} \int_{\epsilon}^{\epsilon+\delta} \exp K \left[-\frac{e^{Kz/q}}{K} + \frac{z}{q}\right] dz \doteq \exp\left[K \max_{y \in [\epsilon, \epsilon+\delta]} \left\{-\frac{e^{Kz/q}}{K} + \frac{z}{q}\right\}\right] \\ &= \exp\left[K \max_{y \in (\epsilon, \epsilon+\delta)} \left\{s_a(y) - \log 2\right\}\right] \end{aligned}$$

The entropy function for the Weibull distribution is given by

$$s_a(y) = \log 2 - \frac{e^{Ky/q}}{K} + \frac{y}{q}, \; y \in [y_l, y_u] \tag{C.4}$$

The bounds of the interval in which the entropy is defined are obtained by setting $s_a(y) = 0$.

Even though an exact formula for the zeros of $s_a(y)$ cannot be obtained, we can make a perturbative analysis to get their asymptotic behavior by balancing the two dominant terms $s_a(y)$ as $M \to \infty$ and analyzing the residual.

Let's assume a balance of the form

$$\log 2 - \frac{\exp\left[\frac{K\widehat{y}_u}{q}\right]}{K} = 0 \Leftrightarrow \widehat{y}_u = \frac{q \log\left(K \log 2\right)}{K} \tag{C.5}$$
$$\mathrm{res}_{\widehat{y}_u}(K) = \frac{\widehat{y}_u}{q} = \frac{\log\left(K \log 2\right)}{K} \underset{K\to\infty}{\longrightarrow} 0$$

The upper zero of the entropy function behaves as $\widehat{y}_u$ and the error produced by this estimation tends to zero as a function of the sample size $\mathrm{res}_{\widehat{y}_u}(M) \approx \dfrac{\log\log M}{\log M}$.
Another possible balance between terms of the equation could be

$$\log 2 + \frac{y_l}{q} = 0 \Leftrightarrow y_l = -q \log 2 \tag{C.6}$$
$$\mathrm{res}_{\widehat{y}_l}(K) = -\frac{1}{K} \exp\left[-K \log 2\right] \underset{K\to\infty}{\longrightarrow} 0$$

In this balance the estimate y_l does not depend on the sample size. Note the residual tends to zero much faster than in previous case, $\mathrm{res}_{\widehat{y}_l}(M) \approx \dfrac{1}{M \log M}$.

An important observation is that $\widehat{y}_l$ underestimates y_l and $\widehat{y}_u$ overestimates y_u (their residuals are negative and positive respectively). So the true range $[y_l, y_u]$ is always contained in $[\widehat{y}_l, \widehat{y}_u]$.
Once we have made this analysis we can compute

$$Z_M \doteq \int_{\widehat{y}_l}^{\widehat{y}_u} \exp\left[K\left(s_a(y) + y\right)\right] dy \doteq \exp\left[K \max_{y\in[\widehat{y}_l,\widehat{y}_u]} \left\{s_a(y) + y\right\}\right] \tag{C.7}$$

For sufficiently large size, the maximum of the exponent is reached at $y^* = \frac{q\log(q+1)}{K}$ and

$$Z_M \doteq \exp\left[-(q+1)\right](q+1)^{q+1} \tag{C.8}$$

as we had estimated by saddle point.

If $y^* \notin [y_l, y_u]$, the maximum of the exponent is reached at the upper bound of the interval y_u. As we mentioned before, y_u can only be computed *exactly* by numerical methods but we can make use the approximation in the perturbative analysis.

$$Z_M \doteq \exp\left[-\exp\left[\frac{K y_u}{q}\right] + K y_u \left(\frac{q+1}{q}\right)\right] \underset{y_u\sim\widehat{y}_u}{\approx} \frac{\left(\log 2^K\right)^{q+1}}{2^K} = \frac{\left(\log M\right)^{q+1}}{M} \tag{C.9}$$

We can also estimate a critical sample size M_c between the two regimes by

$$\widehat{y}_u = y^* \Leftrightarrow \widehat{K}_c = \frac{q+1}{\log 2} \Leftrightarrow \widehat{M}_c = \exp\left[q+1\right] \tag{C.10}$$

Joining all the analysis we can describe the behavior of Z_M,

$$Z_M \doteq \begin{cases} \exp\left[-(q+1)\right](q+1)^{q+1}, \ \lambda \geq 1 \\ \exp\left[-\exp\left[\frac{Ky_u}{q}\right] + Ky_u\left(\frac{q+1}{q}\right)\right], \ \lambda \leq 1 \end{cases} \quad \underset{y_u \approx \widehat{y}_u}{\approx} \tag{C.11}$$

$$\approx \begin{cases} \exp\left[-(q+1)\right](q+1)^{q+1}, \ \widehat{\lambda} \geq 1 \\ \frac{(\log M)^{q+1}}{M}, \ \widehat{\lambda} \leq 1 \end{cases} \tag{C.12}$$

where $\widehat{\lambda}(M,q) = \dfrac{\log M}{q+1}$.

Appendix D

Chi-Squared distribution

The pdf of a Chi-Squared random variable X is

$$f_X(x) = \frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} \exp\left[-x/2\right], d \in \mathbb{N}, x \in [0, \infty) \tag{D.1}$$

where d are the degrees of freedom of the distribution. We note $X \sim \chi(d)$. Chi-squared random variables are additive, that is if $X_1, ..., X_M$ are distributed as $X_i \sim \chi(d_i)$ then the random variable $H = X_1 + ... + X_M \sim \chi(d_1 + ... + d_M)$.

One can express the pdf of a random variable $Y = \log X$ as

$$f_Y(y) = \mathrm{C}_d \exp\left[\frac{-e^y}{2} + \frac{d}{2}y\right], y \in (-\infty, \infty) \tag{D.2}$$

A saddle point estimate of the q-th moment can be given by

$$\mathbb{E}[e^{qY}] = \int_{-\infty}^{\infty} \mathrm{C}_d \exp\left[-\frac{e^y}{2} + \left(\frac{d}{2} + q\right) y\right] dy \doteq \exp\left[\max_{y \in (-\infty, \infty)} \left\{-\frac{e^y}{2} + \left(\frac{d}{2} + q\right) y\right\}\right] =$$

$$= \exp\left[-\left(\frac{d}{2} + q\right) + \left(\frac{d}{2} + q\right) \log\left(d + 2q\right)\right] \tag{D.3}$$

The probability of a sample of being inside an interval $[K\epsilon, K(\epsilon + \delta)]$ is

$$\mathbb{P}\left(Y \in [K\epsilon, K(\epsilon + \delta)]\right) = \int_{K\epsilon}^{K(\epsilon+\delta)} \mathrm{C}_d \exp\left[-\frac{e^y}{2} + \frac{d}{2}y\right] dy$$
$$= \int_{\epsilon}^{\epsilon+\delta} \mathrm{C}_d K \exp\left[K\left(\frac{-e^{Kz}}{2K} + \frac{d}{2}z\right)\right] dz \doteq \exp\left[K \max_{y \in [\epsilon, \epsilon+\delta]} \left\{\frac{-e^{Ky}}{2K} + \frac{d}{2}y\right\}\right] \doteq$$
$$\doteq \exp\left[K \max_{y \in [\epsilon, \epsilon+\delta]} \left\{s_a(y) - \log 2\right\}\right]$$

The entropy function for Chi-Squared distribution is

$$s_a(y) = \log 2 - \frac{e^{Ky}}{2K} + \frac{d}{2}y, y \in [y_l, y_u] \tag{D.4}$$

A perturbative analysis to solve $s_a(y) = 0$ is carried out by balancing the two dominant terms as $M \to \infty$,

$$\log 2 - \frac{e^{K\widehat{y}_u}}{2K} = 0 \Leftrightarrow \widehat{y}_u = \frac{\log\left(2K\log 2\right)}{K} \tag{D.5}$$

$$\mathrm{res}_{\widehat{y}_u} = \frac{d}{2}\,\frac{\log\left(2K\log 2\right)}{K} \underset{K\to\infty}{\longrightarrow} 0 \tag{D.6}$$

where the residual tends to zero like $\dfrac{\log\log M}{\log M}$.

On the other hand,

$$\log 2 + \frac{d}{2}\widehat{y}_l = 0 \Leftrightarrow \widehat{y}_l = -\frac{2}{d}\log 2 \tag{D.7}$$

$$\mathrm{res}_{\widehat{y}_l} = -\frac{\exp\left[-\frac{2K\log 2}{d}\right]}{2K} \underset{K\to\infty}{\longrightarrow} 0 \tag{D.8}$$

and in this case the residual tends to zero much faster than in the previous case, as $\dfrac{1}{M\log M}$.

$$\begin{aligned} Z_M^{(q)} &\approx \frac{1}{M}\int_{y_l}^{y_u} \exp\left[s_a(y) + qy\right] dy \doteq \exp\left[K \max_{y\in[y_l,y_u]}\left\{s_a(y) + qy\right\}\right] = \\ &= \exp\left[K \max_{y\in[y_l,y_u]}\left\{\log 2 - \frac{e^{Ky}}{2K} + \left(\frac{d}{2}+q\right)y\right\}\right] \end{aligned} \tag{D.9}$$

For a sufficient sample size,the maximum is reached at $y^* = \dfrac{\log\left(d+2q\right)}{K}$ and

$$Z_M^{(q)} \doteq \exp\left[-\left(\frac{d}{2}+q\right) + \left(\frac{d}{2}+q\right)\log\left(d+2q\right)\right] \tag{D.10}$$

Note this estimate is exactly the same as saddle point one.

For small sizes, $y^* \notin [y_l, y_u]$ and the maximum is reached at y_u. Evaluating,

$$Z_M^{(q)} \doteq \exp\left[Kqy_u - K\log 2\right] \underset{y_u\approx\widehat{y}_u}{\approx} \exp\left[-K\log 2 + \log\left(2K\log 2\right)\left(\frac{d}{2}+q\right)\right] \tag{D.11}$$

One can set a critical size M_c imposing the border condition,

$$\begin{aligned} y^* = y_u &\Leftrightarrow \frac{\log\left(d+2q\right)}{K_c} = y_u \\ y^* = \widehat{y}_u &\Leftrightarrow \widehat{K}_c = \frac{d+2q}{2\log 2} \Leftrightarrow \widehat{M_c} = \exp\left[\frac{d+2q}{2}\right] \end{aligned} \tag{D.12}$$

Joining all the previous results, the phase transition is described as

$$Z_M^{(q)} \doteq \begin{cases} \exp\left[-\left(\frac{d}{2}+q\right) + \left(\frac{d}{2}+q\right)\log\left(d+2q\right)\right], & \lambda \geq 1 \\ \exp\left[\frac{e^{Ky_u}}{2} + K\left(\frac{d}{2}+q\right)y_u\right], & \lambda \leq 1 \end{cases} \underset{y_u\approx\widehat{y}_u}{\approx} \tag{D.13}$$

$$\approx \begin{cases} \exp\left[-\left(\frac{d}{2}+q\right) + \left(\frac{d}{2}+q\right)\log\left(d+2q\right)\right], & \widehat{\lambda} \geq 1 \\ \exp\left[-K\log 2 + \log\left(2K\log 2\right)\left(\frac{d}{2}+q\right)\right], & \widehat{\lambda} \leq 1 \end{cases} \tag{D.14}$$

where $\widehat{\lambda}(M, d, q) = \dfrac{2 \log M}{d + 2q}$ is the function that controls the transition.

Appendix E

Asymptotic evaluation of integrals

The theory of asymptotic evaluation of integrals can be placed in a more general context of complex analysis. In particular, the study of holomorphic functions, differentiable complex-valued functions, provides theorems to easily compute integrals on complex contours. One of the most important theorems in complex analysis is Cauchy's integral theorem. Basically, it states that the integral of an holomorphic function over a closed contour is always zero. A direct application of this theorem is that it doesn't matter the path we choose, the integral of a holomorphic function between two points is invariant.

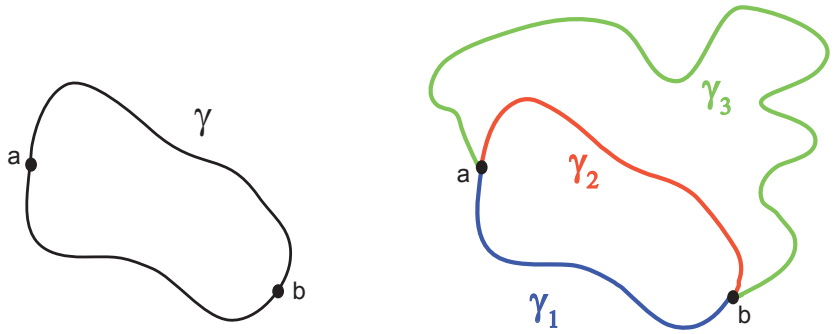


Figure E.1: Illustration of the Cauchy's Integral Theorem. The union of paths γ_1 and γ_2 is exactly γ so the sum of the integrals over them has to be 0 which implies they are equal (orientation of the paths changes the sign of the integral). Applying the same reasoning, the integral over γ_3 is equal to the previous ones.

Given a complex-valued function $f(z) = u(z) + iv(z)$, we define the integral

$$I(K) = \int_a^b g(z) \exp\left[K\ f(z)\right] dz, K >> 0 \tag{E.1}$$

Let's assume f has an only maximum and it is analytic in the whole complex plane. Our intention is to modify the integration countour in order to pass through the maximum (saddle

point) z_0. We can approximate the function in a neighborhood of z_0,

$$f(z) \approx f(z_0) + \frac{1}{2} f''(z_0)(z - z_0)^2$$

and $I(K)$ becomes

$$I(K) = g(z_0) \exp\left[Kf(z_0)\right] \int \exp\left[\frac{K}{2} f''(z_0)(z - z_0)^2\right] dz \tag{E.2}$$

We can parametrize the integrand using the exponential form for complex numbers,

$$f''(z_0) = |f''(z_0)|e^{i\theta}, z - z_0 = re^{i\phi}$$
$$I(K) \approx g(z_0) \exp\left[Kf(z_0)\right] \int \exp\left[\frac{K}{2} |f''(z_0)| e^{i(\theta+2\phi)} r^2\right] e^{i\theta} dr$$

Note θ is fixed, r is the integration variable and we can choose ϕ. Just by imposing $\theta + 2\phi = \pi$ and extending the interval of integration from $(-\infty, \infty)$,

$$I(K) \approx g(z_0) \exp\left[Kf(z_0) + i\theta\right] \int_{-\infty}^{\infty} \exp\left[-\frac{K}{2} |f''(z_0)| r^2\right] dr \tag{E.3}$$

Finally applying the integration of a general Gaussian function

$$\int_{-\infty}^{\infty} e^{-(x+b)^2/c^2} dx = c\sqrt{\pi} \tag{E.4}$$

we get the saddle point estimation of the integral,

$$I(K) \approx \exp\left[Kf(z_0) + i\theta\right] \left(\frac{2\pi}{K|f''(z_0)|}\right)^{1/2} \tag{E.5}$$

The Laplace's method and the stationary phase one, that have been widely used through this document, are particular cases of this general theory taking different values for ϕ ($\phi = 0$ in Laplace's method for considering only real functions and $\theta + 2\phi = \pi/2$ in the stationary phase method).

Appendix F

Monotonic decrease in the bias of the logarithm of the empirical estimator of the mean

The proof given in this section can be considered a generalization of the Jensen's inequality and it is constructed in [36]. First of all, we state this useful result.

Proposition F.0.2 *Given a random variable X and g a convex function, then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)] \tag{F.1}$$

Let's assume g is a increasing function, then g^{-1} is properly defined,

$$\mathbb{E}[X] \leq g^{-1}\left(\mathbb{E}[g(X)]\right) \tag{F.2}$$

The left part of the inequality is a linear average whereas right part will be non-linear in general.
We also define the sample expression of the right part of the inequality,

$$\langle X \rangle_g^M = \mathbb{E}\left[g^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} g(X_i)\right)\right] = \int g^{-1}\left(\frac{g(x_1)+...+g(x_M)}{M}\right)\prod_{i=1}^{M} f_X(x_i)dx_i \tag{F.3}$$

Now we can state the theorem we want to focus on.

Theorem F.0.3 *Consider a sequence $\{X_i\}_{i=1}^M$ of random variables. Then*

$$\langle X \rangle_{M+1}^g \geq \langle X \rangle_M^g \tag{F.4}$$

Firstly, we can decrease the problem complexity to single-sample non-linear average of arbitrary size M. Note we can consider that every n-sample is weighted by its probability $f_{X,M}(\{x_i\}) = \prod_{i=1}^M f_X(x_i)$ and we can construct samples of size $M-1$ by removing one element from the sample (we have M choices to remove) uniformly distributed from the M sized

sample. Let's name $\{\hat{x}_i\}$ a sample of size $M-1$, then

$$f_{X,M-1}(\{\hat{x}_i\}) = \frac{1}{M}\left(\int f_{X,M}\left(x,\hat{x}_1,...,\hat{x}_{M-1}\right)dx + \int f_{X,M}\left(\hat{x}_1,x,...,\hat{x}_{M-1}\right)dx + \right.$$
$$\left. +... + \int f_{X,M}\left(\hat{x}_1,\hat{x}_2,...,x\right)dx\right) = \prod_{i=1}^{M-1} f_X(\hat{x}_i)$$

so the construction is consistent.

Let's name

$$u_M(\{x_i\}) = g^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} g(x_i)\right) \tag{F.5}$$

Just by definition $\mathbb{E}\left[u_M(\{x_i\})\right] = \langle X\rangle_M^g$. We can write the recently defined u_M as

$$u_M(\{x_i\}) = g^{-1}\left(\frac{1}{M}\sum_{j=1}^{M}\frac{1}{M-1}\sum_{i\neq j} g(x_i)\right) \tag{F.6}$$

because every term i in the second summation appears $M-1$ times (in every $j\neq i$). Let's name $\{x_i\}_{[j]}$ the previous sample where we have removed the jth element. Rewriting u_M,

$$u_M(\{x_i\}) = g^{-1}\left(\frac{1}{M}\sum_{j=1}^{M} g(u_{M-1}(\{x_i\}_{[j]}))\right) \tag{F.7}$$

because of definition of u_{M-1} in (F.5). Observing $u_{M-1}(\{x_i\}_{[j]})$ can be considered a random variable of a discrete probability space of size M and applying the inequality (F.2),

$$u_M(\{x_i\}) = g^{-1}\left(\mathbb{E}\left[g\left(u_{M-1}(\{x_i\}_{[j]})\right)\right]\right) \geq \mathbb{E}\left[u_{M-1}(\{x_i\}_{[j]})\right] = \frac{1}{M}\sum_{j=1}^{M} u_{M-1}(\{x_i\}_{[j]}) \tag{F.8}$$

Taking the expectation on the random sequence $\{x_i\}_{i=1}^M$ at both sides of the inequality and using the observation about the construction of the samples of size $M-1$ from one of size M,

$$\langle X\rangle_M^g = \mathbb{E}_{\{x_i\}}\left[u_M(\{x_i\})\right] \geq \mathbb{E}_{\{x_i\}}\left[\frac{1}{M}\sum_{i=1}^{M} u_{M-1}(\{x_i\}_{[j]})\right] =$$
$$= \frac{1}{M}\left[\int g^{-1}\left(\frac{g(x_2)+...+g(x_M)}{M-1}\right) f_{X,M}(x,x_2,...,x_M)dx\ dx_2\ ...\ dx_M + ...\right.$$
$$\left. ... + \int g^{-1}\left(\frac{g(x_1)+...+g(x_{M-1})}{M-1}\right) f_{X,M}(x_1,...,x_{M-1},x)dx_1...dx_{M-1}dx_M\right] =$$
$$= \frac{1}{M}\left[M\int g^{-1}\left(\frac{x_1+...+x_{M-1}}{M-1}\right)\prod_{i=1}^{M-1} f_X(x_i)dx_i\right] = \langle X\rangle_{M-1}^g$$

where we have used the factorizability of the probability density function. So we have obtained $\langle X\rangle_M^g \geq \langle X\rangle_{M-1}^g$.

The proof of the monotonicity of the logarithm bias is just an application of this theorem just by taking $g(x) = e^x$.

Bibliography

[1] M. Ljungberg, S. Strand, and M. King, *Monte Carlo calculations in nuclear medicine: applications in diagnostic imaging.* Institute of Physics Pub., 1998.

[2] A. Dubi, *Monte Carlo applications in systems engineering.* John Wiley & Sons, 2000.

[3] P. Glasserman, *Monte Carlo Methods in Financial Engineering (Stochastic Modelling and Applied Probability)*, 1st ed. Springer, 2003.

[4] P. Jäckel, *Monte Carlo methods in finance.* J. Wiley, 2002.

[5] A. S. L. Zhenquin, "Monte carlo minimization approach to the multiple-minima problem in protein folding," *Proc. National Academy of Sciences*, vol. 84, pp. 6611–6615, 1987.

[6] A. R. Leach, *Molecular modelling: principles and applications.* Prentice Hall, 2001.

[7] A. M. Ferrenberg and R. H. Swendsen, "New monte carlo technique for studying phase transitions," *Phys. Rev. Lett.*, vol. 61, pp. 2635–2638, Dec 1988.

[8] W. Gilks, W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice.* Chapman & Hall, 1996.

[9] M. E. Crovella and L. Lipsky, "Long-lasting transient conditions in simulations with heavy-tailed workloads," 1997. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.155.483

[10] B. Derrida, "Random-energy model: An exactly solvable model of disordered systems," *Phys. Rev. B*, vol. 24, pp. 2613–2626, Sep 1981.

[11] H. B. Mann and A. Wald, "On stochastic limit and order relationships." 1943.

[12] A. Kagan and S. Nagaev, "How many moments can be estimated from a large sample?" *Statistics & Probability Letters*, vol. 55, no. 1, pp. 99–105, 2001.

[13] M. Mézard and A. Montanari, *Information, physics, and computation.* Oxford University Press, 2009.

[14] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling extremal events for insurance and finance.* Springer, 1997.

[15] H. Touchette, "The large deviation approach to statistical mechanics," *Physics Reports*, vol. 478, no. 1-3, pp. 1–69, Aug. 2009.

[16] J. W. T.C. Dorlas, "Large deviations theory and the random energy model," *International Journal of Modern Physics B*, vol. 15, pp. 1–15, 2001.

[17] C. Bender and S. Orszag, *Advanced mathematical methods for scientists and engineers: Asymptotic methods and perturbation theory.* Springer, 1978.

[18] N. K. Jana, "Contributions to Random Energy Models," Nov. 2007. [Online]. Available: http://arxiv.org/abs/0711.1249

[19] F. Angeletti, E. Bertin, and P. Abry, "Critical moment definition and estimation, for finite size observation of log-exponential-power law random variables," Mar. 2011. [Online]. Available: http://arxiv.org/abs/1103.5033

[20] J. Laherrere and D. Sornette, "Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales," *The European Physical Journal B*, vol. 2, p. 525, 1998.

[21] C. G. Justus, W. R. Hargraves, A. Mikhail, and D. Graber, "Methods for estimating wind speed frequency distributions," *Journal of Applied Meteorology*, vol. 17, no. 3, pp. 350–353, 1978.

[22] R. Elandt-Johnson and N. Johnson, *Survival Models and Data Analysis.* Wiley, 1980.

[23] A. Mood, F. Graybill, and D. Boes, *Introduction to the theory of statistics.* McGraw-Hill, 1974.

[24] A.-L. Barabási and E. Bonabeau, "Scale-free networks." *Scientific American*, vol. 288, no. 5, pp. 60–69, 2003.

[25] "Random texts exhibit zipf's-law-like word frequency distribution," *IEEE Transactions on Information Theory*, vol. 38, pp. 1842–1845, 1992.

[26] M. S. Mega, P. Allegrini, P. Grigolini, V. Latora, L. Palatella, A. Rapisarda, and S. Vinciguerra, "Power-law time distribution of large earthquakes," *Phys. Rev. Lett.*, vol. 90, p. 188501, May 2003.

[27] L. T. R.A. Fisher, "Limiting forms of the frequency distribution of the largest and smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 180–190, 1927.

[28] B. Gnedenko, "Sur la distribution limite du terme maximum d'une serie aleatoire." *The Annals of Mathematics*, vol. 44, no. 3, pp. 423–453, 1943.

[29] M. L. A. Bovier, I. Kurkova, "Fluctuations of the free energy in the REM and the p-spin SK models," *Annals of Probability*, vol. 30, no. 2, pp. 605–651, 2002.

[30] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions.* Springer, 2011.

[31] C. Klupperlberg, "Subexponential distributions and integrated tails," *J. Appl. Prob*, vol. 25, no. 1, pp. 132–141, 1988.

[32] D. L. Wallace, "Asymptotic approximations to distributions," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 635–654, 1958.

[33] G. Stüber, *Principles of Mobile Communication*. Springer, 2011.

[34] J. J. Wu, N. Mehta, "A flexible lognormal sum approximation method," *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 6, pp. 3413 – 3417, 2005.

[35] L. N. M. M. H. S. P. Gopikrishnan, V. Plerou, "Scaling of the distribution of fluctuations in financial markets," *Phys. Rev. E 60*, pp. 5305–5316, 1999.

[36] D. Zuckerman and T. Woolf, "Systematic finite-sampling inaccuracy in free energy differences and other nonlinear quantities," *Journal of Statistical Physics*, vol. 114, pp. 1303–1323, 2004.