On the Interplay of Deviation Inequalities: Berry-Esseen, Chernoff, and Moderate Deviations

Gemini

October 11, 2025

Abstract

The Central Limit Theorem (CLT) is a cornerstone of probability theory, describing the convergence of sums of random variables to a normal distribution. However, the CLT is a qualitative statement. This paper explores three fundamental results that provide a quantitative understanding of deviations from the mean across a spectrum of scales. We examine the Berry-Esseen theorem, which quantifies the rate of convergence in the CLT for typical deviations; the Chernoff inequality and the associated theory of large deviations, which provide exponential bounds for rare, large-scale events; and the Moderate Deviation Principle, which elegantly bridges the gap between these two regimes. By analyzing their respective domains of applicability and the nature of their bounds, we present a unified perspective on how these principles collectively offer a comprehensive picture of the concentration of measure phenomenon.

1 Introduction

The study of sums of independent random variables is a central theme in probability theory and statistics. The Law of Large Numbers (LLN) asserts that the sample mean of a sequence of independent and identically distributed (i.i.d.) random variables converges to the true mean. The Central Limit Theorem (CLT) goes further, describing the fluctuations around this mean. It states that the standardized sum of these variables converges in distribution to a standard normal distribution.

While powerful, the classical CLT is a limiting result; it does not specify the probability of deviations for a finite number of variables, nor does it adequately describe the probability of very large, or ”rare,” deviations. This paper delves into the theoretical tools designed to address these quantitative questions. We aim to elucidate the relationships between three key frameworks:

1. **The Berry-Esseen Theorem**, [1, 4] which provides a non-asymptotic bound on the error of the normal approximation, thereby quantifying the CLT for typical deviations.

2. **The Chernoff Inequality**, a technique that yields exponential upper bounds on the tail probabilities of sums, forming the basis of the theory of large deviations.

3. **The Moderate Deviation Principle (MDP)**, a more recent development that characterizes deviation probabilities on a scale intermediate between the CLT and large deviations.

By examining these three principles in concert, we reveal a spectrum of deviation behaviors, from the Gaussian fluctuations of order $O(\sqrt{n})$ to the rare events of order $O(n)$. This exploration demonstrates that these are not competing theories but complementary components of a single, rich story about how sums of random variables concentrate around their mean.

2 Preliminaries and Notation

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables. We define the following:

- Mean: $\mathbb{E}[X_i] = \mu$
- Variance: $\mathrm{Var}(X_i) = \mathbb{E}[(X_i - \mu)^2] = \sigma^2 \in (0, \infty)$
- Partial Sum: $S_n = \sum_{i=1}^{n} X_i$
- Standardized Sum: $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$

The cumulative distribution function (CDF) of the standard normal distribution is denoted by $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$.

Principle 2.1 (Central Limit Theorem)**.** *As $n \to \infty$, the CDF of the standardized sum Z_n converges to the standard normal CDF:*

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad \forall x \in \mathbb{R}$$

For our discussion of Chernoff bounds, we will also require the moment-generating function (MGF) of a random variable X, defined as $M_X(t) = \mathbb{E}[e^{tX}]$, and the related cumulant generating function (CGF), $\Lambda_X(t) = \log M_X(t)$.

3 The Berry-Esseen Theorem: A Quantitative CLT

The Berry-Esseen theorem answers the question: how fast does the distribution of Z_n converge to $\Phi(x)$? It provides a uniform, non-asymptotic bound on the approximation error.

Theorem 3.1 (Berry-Esseen)**.** *Assume that $\mathbb{E}[|X_1|^3] < \infty$. Let $\rho = \mathbb{E}[|X_1 - \mu|^3]$ be the third absolute central moment. Then there exists a universal constant $C > 0$ such that for all $n \geq 1$:*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x \right) - \Phi(x) \right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

3.1 Interpretation and Significance

The crucial insight from the Berry-Esseen theorem is that the rate of convergence in the CLT is of the order $O(n^{-1/2})$. This result transforms the CLT from an asymptotic statement into a practical tool for finite n, providing an explicit error bound.

The theorem's domain is the scale of "typical" deviations from the mean—fluctuations of magnitude proportional to $\sqrt{n}$. It guarantees that within this central region of the distribution, the normal approximation is accurate up to a quantifiable, power-law error term.

3.2 Limitations

The Berry-Esseen theorem's strength in the central part of the distribution is also its limitation. The uniform bound it provides is often not informative for the tails of the distribution. For a fixed large x, both $\mathbb{P}(Z_n > x)$ and $1 - \Phi(x)$ may be much smaller than the error bound $O(n^{-1/2})$. Therefore, it fails to capture the rate of decay of tail probabilities, which is often exponential. To analyze these rare events, a different set of tools is required.

4 Chernoff Bounds and the Theory of Large Deviations

The theory of large deviations is concerned with the asymptotic probability of rare events [2, 3]. The Chernoff bound is the foundational technique in this area.

4.1 The Chernoff Bounding Technique

The technique relies on applying Markov's inequality to the MGF of the sum. For any $a \in \mathbb{R}$ and any $t > 0$:

$$\mathbb{P}(S_n \geq a) = \mathbb{P}(e^{tS_n} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tS_n}]}{e^{ta}}$$

Since the variables are i.i.d., $\mathbb{E}[e^{tS_n}] = \left(\mathbb{E}[e^{tX_1}]\right)^n$. The bound holds for any $t > 0$, so we can optimize it:

$$\mathbb{P}(S_n \geq a) \leq \inf_{t>0} e^{-ta} \left(M_{X_1}(t)\right)^n$$

4.2 Cramér's Theorem and the Rate Function

This technique is formalized by Cramér's theorem, a cornerstone of Large Deviation Theory. Without loss of generality, assume $\mu = 0$.

Theorem 4.1 (Cramér's Theorem)**.** *Let S_n be the sum of i.i.d. mean-zero random variables. The probabilities $\mathbb{P}(S_n/n \geq x)$ for $x > 0$ decay exponentially with a rate governed by a "rate function" $I(x)$:*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \geq x\right) = -I(x)$$

where the rate function $I(x)$ is the Legendre-Fenchel transform of the CGF:

$$I(x) = \sup_{t\in\mathbb{R}} \{tx - \Lambda_{X_1}(t)\}$$

This states that for large n, $\mathbb{P}(S_n/n \geq x) \approx e^{-nI(x)}$.

4.3 Interpretation and Contrast with Berry-Esseen

Large Deviation Theory addresses deviations of order $O(n)$, in stark contrast to the $O(\sqrt{n})$ scale of the CLT. The probability of such an event is exponentially small in n. While Berry-Esseen provides a power-law ($n^{-1/2}$) correction to the probability density in the central region, Chernoff bounds provide the exponential rate of decay in the tails. They are fundamentally different tools for different regimes.

5 The Moderate Deviation Principle: Bridging the Gap

A natural question arises: what happens for deviations that are larger than the CLT scale but smaller than the large deviation scale? This is the regime of "moderate" deviations.

Principle 5.1 (Moderate Deviation Principle (MDP))**.** *Let S_n be a sum of i.i.d. random variables with $\mu = 0$ and $\sigma^2 = 1$. Let $\{a_n\}_{n\geq 1}$ be a sequence of positive numbers such that:*

$$a_n \to \infty \quad \text{and} \quad \frac{a_n}{\sqrt{n}} \to 0 \quad \text{as} \quad n \to \infty$$

Then, for any fixed $x > 0$, the sequence of random variables $\frac{S_n}{a_n\sqrt{n}}$ satisfies a large deviation principle with speed a_n^2 and rate function $I_0(x) = x^2/2$. That is:

$$\lim_{n\to\infty} \frac{1}{a_n^2} \log \mathbb{P}\left(\frac{S_n}{a_n\sqrt{n}} \geq x\right) = -\frac{x^2}{2}$$

This implies that for large n, $\mathbb{P}(S_n \geq x a_n \sqrt{n}) \approx \exp(-a_n^2 x^2/2)$.

5.1 The Full Spectrum of Deviations

The MDP beautifully connects the worlds of the CLT and large deviations. Consider the probability of the event $S_n \geq y_n$. The behavior depends on the growth rate of the threshold y_n:

- **CLT Scale:** If $y_n = O(\sqrt{n})$, say $y_n = x\sigma\sqrt{n}$, then $\mathbb{P}(S_n \geq y_n) \to 1 - \Phi(x)$. The error in this approximation is $O(n^{-1/2})$ by the Berry-Esseen theorem.

- **Moderate Deviation Scale:** If $y_n = x a_n \sigma \sqrt{n}$ where a_n satisfies the MDP conditions, the probability decays as $\exp(-a_n^2 x^2/2)$. This is faster than any power law but slower than the exponential decay of large deviations.

- **Large Deviation Scale:** If $y_n = O(n)$, say $y_n = xn$, the probability decays as $\exp(-nI(x))$.

The quadratic rate function $I_0(x) = x^2/2$ of the MDP is not arbitrary. It is precisely the second-order Taylor approximation of the large deviation rate function $I(x)$ around its minimum at $x = 0$. This shows that moderate deviations are those large enough to escape the Gaussian core but not so large as to feel the full, variable curvature of the large deviation rate function.

6 Conclusion

The Berry-Esseen theorem, Chernoff's inequality, and the Moderate Deviation Principle provide a multi-layered understanding of the concentration of sums of random variables. They are not isolated results but rather points on a continuous spectrum of deviation analysis.

- **Berry-Esseen** gives the precise power-law error for the normal approximation in the $O(\sqrt{n})$ central region.
- **Chernoff bounds** and large deviation theory provide the exponential decay rate for rare events in the $O(n)$ tail region.
- **Moderate Deviations** seamlessly connect these two regimes, showing how the Gaussian-like quadratic rate function emerges as the first-order behavior before the full large deviation rate function takes over for even larger deviations.

Together, these three frameworks offer a powerful and remarkably coherent toolkit for probabilists and statisticians, enabling a deep and quantitative analysis of one of mathematics' most fundamental objects: the sum of independent random variables.

7 MDP for Eigenvalue Counts in Random Matrices

Random Matrix Theory (RMT) provides a powerful framework for modeling complex systems where the exact Hamiltonian is unknown. A central focus of RMT is the statistical behavior of eigenvalues. It is well-established that eigenvalue fluctuations in the "bulk" of the spectrum are governed by a Central Limit Theorem (CLT), while those at the spectral "edge" are described by the Tracy-Widom distribution. This paper explores the transitional regime between these two distinct behaviors. We demonstrate that a Moderate Deviation Principle (MDP) precisely characterizes the tails of the distribution for eigenvalue counts in a mesoscopic window, providing a unified mathematical description that smoothly connects the Gaussian fluctuations of the bulk to the non-classical behavior at the edge.

In many complex quantum and statistical systems, from heavy atomic nuclei to financial markets, the statistical properties of energy levels or characteristic modes can be effectively modeled by the eigenvalues of large random matrices [5, 6]. For an $N \times N$ random matrix drawn from a suitable ensemble, the eigenvalues $\{\lambda_i\}_{i=1}^N$ exhibit remarkable universal properties as $N \to \infty$.

The study of these eigenvalues has revealed two fundamentally different fluctuation regimes:

1. **The Bulk:** Deep inside the spectrum, the number of eigenvalues within a fixed interval, after appropriate centering and scaling, converges to a Gaussian distribution. This is a manifestation of a Central Limit Theorem (CLT).
2. **The Edge:** At the extreme ends of the spectrum, the largest (or smallest) eigenvalue exhibits fluctuations described by the Tracy-Widom distribution [7], a hallmark of universality classes in growth models and other stochastic systems.

A natural and important question arises: how does the system transition from the Gaussian behavior of the bulk to the Tracy-Widom behavior at the edge? Standard CLT and Large Deviation Principle (LDP) frameworks are insufficient to describe this crossover. The CLT only captures typical fluctuations, while the LDP describes exponentially rare events. The transition occurs at an intermediate, or ”mesoscopic,” scale.

This paper formally demonstrates that the answer lies in the theory of **Moderate Deviations**. A Moderate Deviation Principle (MDP) provides a precise asymptotic for the probabilities of deviations that are larger than the CLT scale but smaller than the LDP scale. We will show how the MDP for linear spectral statistics (i.e., eigenvalue counts in a window) guides the tails of the count distribution, providing a continuous bridge between the bulk and edge regimes.

8 Background and Definitions

To formalize our discussion, we introduce the necessary concepts from Random Matrix Theory and probability theory.

8.1 The Gaussian Unitary Ensemble (GUE)

The GUE is a canonical example of a random matrix ensemble.

Definition 8.1 (Gaussian Unitary Ensemble (GUE))**.** The GUE(N) is a probability measure on the space of $N \times N$ Hermitian matrices $M = (M_{jk})$ defined by the density:

$$dP_N(M) = \frac{1}{Z_N} \exp\left(-\frac{N}{2}\mathrm{Tr}(M^2)\right) dM$$

where $dM = \prod_{j=1}^{N} dM_{jj} \prod_{1\le j<k\le N} d\mathrm{Re}(M_{jk}) d\mathrm{Im}(M_{jk})$ is the Lebesgue measure on the real and imaginary parts of the entries, and Z_N is a normalization constant.

The joint probability density function (j.p.d.f.) of the real eigenvalues $\lambda_1, \dots, \lambda_N$ of a GUE matrix is given by:

$$p(\lambda_1, \dots, \lambda_N) = \frac{1}{Z'_N} \prod_{1\le j<k\le N} (\lambda_k - \lambda_j)^2 \exp\left(-\frac{N}{2}\sum_{i=1}^{N} \lambda_i^2\right)$$

The term $\prod(\lambda_k - \lambda_j)^2$, known as the Vandermonde determinant squared, encodes the ”repulsion” between eigenvalues, a fundamental feature of RMT.

8.2 Eigenvalue Statistics and the Semicircle Law

We study eigenvalues through their empirical spectral distribution (ESD).

Definition 8.2 (Empirical Spectral Distribution)**.** The ESD of a matrix M with eigenvalues $\{\lambda_i\}_{i=1}^N$ is the probability measure

$$L_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}$$

where δ_x is the Dirac delta mass at x.

For GUE, as $N \to \infty$, the ESD converges weakly to a deterministic measure known as the Wigner semicircle law.

Theorem 8.1 (Wigner's Semicircle Law)**.** *Almost surely, as $N \to \infty$, the ESD L_N of a GUE matrix converges weakly to the semicircle distribution σ_{sc}, with density:*

$$\rho_{sc}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad x \in [-2, 2]$$

and $\rho_{sc}(x) = 0$ otherwise.

The support of this density, $[-2, 2]$, is referred to as the **bulk** of the spectrum, and the points ±2 are the spectral **edges**.

The number of eigenvalues in an interval $I \subset \mathbb{R}$, denoted $\mathcal{N}_I$, can be written as a linear spectral statistic:

$$\mathcal{N}_I = N \int_I dL_N(\lambda) = \sum_{i=1}^{N} \mathbf{1}_I(\lambda_i)$$

where $\mathbf{1}_I$ is the indicator function for the interval I. The semicircle law implies the expectation converges as $\mathbb{E}[\mathcal{N}_I] \approx N \int_I \rho_{sc}(x)dx$.

8.3 Deviation Principles

We formally define the probabilistic tools used to study fluctuations.

Definition 8.3 (Large and Moderate Deviation Principles)**.** Let $\{X_N\}_{N\geq 1}$ be a sequence of random variables.

1. It satisfies a **Large Deviation Principle (LDP)** with speed $v_N \to \infty$ and rate function $J(x)$ if J is a non-negative, lower semi-continuous function and for any Borel set A:

$$-\inf_{x\in A^\circ} J(x) \leq \liminf_{N\to\infty} \frac{1}{v_N} \log \mathbb{P}(X_N \in A) \leq \limsup_{N\to\infty} \frac{1}{v_N} \log \mathbb{P}(X_N \in A) \leq -\inf_{x\in \bar{A}} J(x)$$

   For linear statistics of GUE eigenvalues, the speed is $v_N = N^2$.

2. It satisfies a **Moderate Deviation Principle (MDP)** with speed $a_N \to \infty$ and scaling $b_N \to \infty$ such that $a_N/b_N \to 0$, if the sequence $Y_N = \sqrt{a_N/b_N}(X_N - \mathbb{E}[X_N])$ satisfies an LDP with speed b_N/a_N and some rate function $I(x)$.

Essentially, the MDP examines fluctuations at an intermediate scale between the CLT scale (where $a_N = 1, b_N = 1$) and the LDP scale.

9 Main Result: MDP for Windowed Counts

We now consider the fluctuations of the centered and scaled eigenvalue count $\mathcal{N}_I$. Let $I \subset (-2, 2)$ be a fixed interval in the bulk. The classical CLT for linear statistics states that:

$$\frac{\mathcal{N}_I - \mathbb{E}[\mathcal{N}_I]}{\sqrt{\frac{1}{2\pi^2}\log N}} \xrightarrow{d} \mathcal{N}(0, 1)$$

This describes typical fluctuations of order $\mathcal{O}(\sqrt{\log N})$. The Tracy-Widom law, on the other hand, describes fluctuations of the edge eigenvalues, which are of order $\mathcal{O}(N^{-2/3})$.

The MDP bridges these results by considering deviations of size b_N where $\sqrt{\log N} \ll b_N \ll N$.

Theorem 9.1 (MDP for Eigenvalue Counts, informal)**.** *Let I_N be a sequence of intervals within the bulk of the spectrum. Consider the number of eigenvalues $\mathcal{N}_{I_N}$. Let b_N be a sequence satisfying $\sqrt{\log N} \ll b_N \ll N \int_{I_N} \rho_{sc}(x)dx$. Then the sequence of random variables*

$$X_N = \frac{\mathcal{N}_{I_N} - \mathbb{E}[\mathcal{N}_{I_N}]}{b_N}$$

satisfies a Moderate Deviation Principle with speed $\frac{b_N^2}{\log N}$ and a quadratic (Gaussian) rate function $I(x) = \pi^2 x^2$. That is, for large N, the probability of observing a deviation of size $b_N x$ from the mean behaves as:

$$\mathbb{P}\left(\frac{\mathcal{N}_{I_N} - \mathbb{E}[\mathcal{N}_{I_N}]}{b_N} \approx x\right) \approx \exp\left(-\frac{b_N^2}{\log N} \cdot \pi^2 x^2\right)$$

This theorem is powerful because it holds uniformly as the interval I_N approaches the spectral edge. The MDP framework provides a single formula that correctly captures the tail probabilities of eigenvalue counts across a wide range of mesoscopic scales.

9.1 Guiding the Transition

How does this MDP guide the transition from bulk to edge?

- **Approaching the CLT:** If we choose a small deviation scale $b_N \sim \sqrt{\log N}$, the speed of the MDP becomes $\mathcal{O}(1)$. This recovers the Gaussian density of the CLT, showing that the MDP is consistent with the bulk behavior.

- **Moving Towards the Edge:** As the interval I_N shifts towards the edge (e.g., $I_N = [2 - \epsilon_N, 2]$), the underlying assumptions for the CLT break down. The eigenvalue correlations become stronger and non-universal terms become relevant. The MDP framework, however, remains valid for a wider range of deviation scales b_N. It shows that the probability distribution remains Gaussian in shape for moderately large deviations, but the "variance" of this effective Gaussian (controlled by the speed $b_N^2/\log N$) changes. The principle quantifies the increasing likelihood of large deviations as one moves away from the CLT regime towards the highly correlated edge regime. It marks the precise point where the Gaussian approximation of the CLT fails and larger, non-universal deviations begin to dominate, paving the way for the Tracy-Widom statistics that govern the edge itself.

10 Conclusion

The statistical behavior of eigenvalues in large random matrices is a rich and varied field. While the Central Limit Theorem and the Tracy-Widom distribution perfectly describe the fluctuations in the bulk and at the edge, respectively, they leave a gap in our understanding of the transitional regime. This work demonstrates that the Moderate Deviation Principle fills this gap. By providing an explicit, Gaussian-like form for the probabilities of moderately large fluctuations, the MDP serves as a mathematical guide, charting the smooth but rapid evolution of eigenvalue statistics as one moves from the uncorrelated bulk to the highly correlated spectral edge. This provides a more complete and unified picture of the universal laws governing complex systems.

References

[1] A. C. Berry. The accuracy of the gaussian approximation to the distribution of sums of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

[2] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[3] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer Science & Business Media, 2009.

[4] C. G. Esseen. On the liapounoff limit of error in the theory of probability. *Arkiv f"or matematik, astronomi och fysik*, A28(9):1–19, 1942.

[5] M. L. Mehta. *Random Matrices.* Elsevier, 3rd edition, 2004.

[6] T. Tao. *Topics in Random Matrix Theory.* American Mathematical Society, 2012.

[7] C. A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.