American University in Bulgaria

Department of Mathematics

A Senior Thesis in Mathematics

The Large Deviation Principle

By Giorgi Chavchanidze

Supervisor: Alexander Ganchev, PhD

Blagoevgrad, 2018

Abstract

Probability distributions of many random variables defined on a sample of size n are known to be getting concentrated around a particular value exponentially fast with increasing sample size. The Large Deviation Principle (LDP) is an abstraction and a rigorous description of such behavior. In this paper we explore several statements of the LDP and give a friendly yet rigorous exposition of its essence using the notion of exponential decay and little-o notation. Further, we prove some sufficient conditions under which one could infer that the LDP is satisfied by a random variable just by looking at its density/mass function. We also comment about the relationship between the LDP and the Central Limit Theorem, and prove some sufficient conditions under which the satisfaction of the LDP implies that a Generalized Law of Large Numbers holds.

Keywords: Large Deviation Principle, Exponential Decay, Law of Large Numbers, Central Limit Theorem

Claims of Originality

There are claims of originality at the beginning of the sections 2.3, 3 and 4. Please note that these claims pertain to proofs, not necessarily to results, since my literature review is not very extensive.

Contents

1 Introductory Example and Discussion **3**

2 The Large Deviation Principle **6**

2.1 Domination 6

2.2 Large Deviation Principle and Exponential Decay 8

2.3 LDP In Terms of Strictly Convex Rate Functions and Tails of Probability Distribution 9

2.4 Precise Statement of the Large Deviation Principle 13

2.5 Gärtner-Ellis Theorem and iid Sample Mean 14

3 Inferring Large Deviations From Density/Mass Functions **15**

3.1 The Discrete Case 16

3.2 The Continuous Case 18

4 Weak Law of Large Numbers and the Large Deviation Principle **23**

References **25**

1 Introductory Example and Discussion

Let (X_n) be a sequence of independent and identically distributed (iid) bernoulli trials each with sucess probability p and let $M_n = \frac{\sum_{k=1}^{n} X_k}{n}$ be the ratio of successes in n number of trials. We know that by the Law of Large Numbers (LLN) the probability of M_n assuming values in an arbitrarily small open interval around $E(M_n) = p = E(X)$ goes to 1, i.e. probability distribution of M_n gets concentrated around p. We thus call p a typical value of M_n (which is in line with the essence of mathematical expectation). Further, if convergence in probability to p is exponentially fast, we could informally say that M_n satisfies the Large Deviation Principle. Indeed, we shall see shortly that the probability of any atypical outcome of M_n (i.e. a realization of M_n that deviates from a typical value p) decays exponentially fast (in fact, as we will see later, the same is true for any iid sample mean, for which a typical value coincides with the expectation, as it should). This explains the word "Deviations" in "Large Deviations". Further, the Central Limit Theorem (CLT) tells us that the probability distribution of M_n is asymptotically normall as it gets concentrated around p. However, asymptotic normality holds only near p and the CLT fails to provide good approximation for probabilities of deviations that are of order greater than $n^{-\frac{1}{2}}$ (see [7] for a more detailed discussion). The tools of Large Deviations Theory, however, apply to any deviation no matter how large and is in this sense an extention of the CLT. Hence "Large" in "Large Deviations". We illustrate these ideas in more detail by computing a probability mass function of M_n. We follow the derivation given in [4].

We want to estimate the probability of a ratio of successes in n number of trials, i.e. $\Pr(M_n = \frac{\lfloor ns \rfloor}{n})$, or equivalently, $\Pr(S_n = \lfloor ns \rfloor)$, where $S_n = \sum_{k=1}^{n} X_k$, $s \in (0,1)$, and $\lfloor ns \rfloor$ denotes the largest integer less than or equal to ns. We then have by binomial distribution that

$$\begin{aligned} P(S_n = \lfloor ns \rfloor) &= \binom{n}{\lfloor ns \rfloor} p^{\lfloor ns \rfloor}(1-p)^{n-\lfloor ns \rfloor} = \\ &= \frac{n!}{\lfloor ns \rfloor!(n-\lfloor ns \rfloor)!} p^{\lfloor ns \rfloor}(1-p)^{n-\lfloor ns \rfloor} \approx \\ &\approx \frac{n^n p^{\lfloor ns \rfloor}(1-p)^{n-\lfloor ns \rfloor}}{\lfloor ns \rfloor^{\lfloor ns \rfloor}(n-\lfloor ns \rfloor)^{n-\lfloor ns \rfloor}} \sqrt{\frac{n}{2\pi \lfloor ns \rfloor(n-\lfloor ns \rfloor)}}, \end{aligned} \tag{1.1}$$

where the approximation is given by Stirling's formula - $m! \approx e^{-m} m^m \sqrt{2\pi m}$, which

means that

$$\lim_{n\to\infty} \frac{m!}{e^{-m}m^m\sqrt{2\pi m}} = 1.$$

We will discuss the meaning of " $\approx$ " and its relation to little-o notation in greater detail in Section 2.1. For now we will simply state that " $\approx$ " means that in the limit the approximation error is an arbitrarily small fraction of the approximation (we will see shortly that our approximation is a decaying exponential in n; it then follows that in the limit the error is shrinking to 0 at least exponentially fast).

Observe that if $\alpha_n = \frac{n^n p^{\lfloor ns\rfloor}(1-p)^{n-\lfloor ns\rfloor}}{\lfloor ns\rfloor^{\lfloor ns\rfloor}(n-\lfloor ns\rfloor)^{n-\lfloor ns\rfloor}} \times \frac{ns^{ns}(n-ns)^{n-ns}}{n^n p^{ns}(1-p)^{n-ns}}$ and $\beta_n = \sqrt{\frac{n}{2\pi\lfloor ns\rfloor(n-\lfloor ns\rfloor)}}$,

$$P(S_n = \lfloor ns\rfloor) \approx \alpha_n\beta_n \frac{n^n p^{ns}(1-p)^{n-ns}}{ns^{ns}(n-ns)^{n-ns}}.$$

By direct computation, we have that

$$\frac{n^n p^{ns}(1-p)^{n-ns}}{ns^{ns}(n-ns)^{n-ns}} = e^{log(\frac{n^n p^{ns}(1-p)^{n-ns}}{ns^{ns}(n-ns)^{n-ns}})} = e^{-nI_p(s)},$$

where

$$I_p(s) = s\log\frac{s}{p} + (1-s)\log\frac{1-s}{1-p}.$$

Moreover, using the following bounds (all of which follow from the definition of $\lfloor ns\rfloor$),

$$\begin{aligned} ns - 1 < \lfloor ns\rfloor < ns \\ n(1-s) < n - \lfloor ns\rfloor < n(1-s)+1 \\ n^2 s(1-s) - n(1-s) < \lfloor ns\rfloor(n-\lfloor ns\rfloor) < n^2 s(1-s) + ns \end{aligned}$$

it can be shown that $o(n) \le \alpha_n\beta_n \le o(n)$, where $o(n)$ denotes a sublinear expression, i.e. an expression $g(n)$ such that $\frac{g(n)}{n} \to 0$ as $n \to \infty$ (a detailed discussion of the meaning of $o(n)$ follows in Section 2.1). Thus, $\alpha_n\beta_n = o(n)$ and so does $\log(\alpha_n\beta_n)$. Hence, we have that

$$P(S_n = \lfloor ns\rfloor) \approx \alpha_n\beta_n \frac{n^n p^{ns}(1-p)^{n-ns}}{ns^{ns}(n-ns)^{n-ns}} = e^{-nI_p(s)+o(n)}$$

and

$$\lim_{n\to\infty} \frac{\log P(S_n = \lfloor ns\rfloor)}{n} = \frac{-nI_p(s)+o(n)}{n} = -I_p(s),$$

where $\lim_{n\to\infty} \frac{\log P(S_n=\lfloor ns\rfloor)}{n}$ is known as the large deviation limit and $I_p(s)$ is called a rate function of S_n (or equivalently of M_n). Also note that although for $s = 0, 1$

the approximation is not defined, we have by direct computation that $P(S_n = \lfloor ns \rfloor) = \binom{n}{\lfloor ns \rfloor} p^{\lfloor ns \rfloor}(1-p)^{n-\lfloor ns \rfloor} = e^{-nI_p(s)}$.

Observe that $I_p(p) = 0$ and $0 < I_p(s) < \infty$ for all $s \neq p$, $s \in [0,1]$. As we will see in Section 2.2, this is equivalent to the statement that every event of the form $S_n = \lfloor ns \rfloor$ with $s \neq p$ decays exponentially fast with rate $I_p(s)$. The existence of exponential decay usually suggests that the LDP is satisfied, and we could informally state that the LDP holds whenever the large deviation limit exists, is finite-valued and not identically zero (such informal formulation is given, for instance, in [6]). In Sections 2.2 and 2.3, we will give a more rigorous formulation of the LDP and in Section 3 we will revisit our introductory example in light of a more precise formulation. In Section 3, we will also show how one can in some cases infer from probability mass/density functions that a random variable satisfies the precise statement of the LDP.

Note that $I_p'(s) = \log \frac{s}{p} - \log \frac{1-s}{1-p}$, and $I_p''(s) = \frac{1}{s(1-s)}$. Thus, $I_p(s)$ is strictly convex: it increases at a rising pace as s moves away from p. We will see in Section 2.5 that strictly convex rate functions are common, and in Section 2.3 we give a formulation of the LDP in terms of strictly convex rate functions and tails of a probability distribution.

The paragraph that follows is inspired by a comment about the LDP and the CLT in [2]. Observe that $I_p^{-1}(0) = p = E(X)$, $I_p'(p) = 0$, and $I_p''(p) = \frac{1}{p(1-p)} = \frac{1}{\sigma_X^2}$. Thus, if we let $s = p + \frac{u}{\sqrt{n}}$ and take a second-order Taylor series expansion of $I_p(s)$ around p, we have that

$$I_p(s) = I_p(p) + I_p'(p)\frac{u}{\sqrt{n}} + \frac{1}{2}I_p''(p)\frac{u^2}{n} + O(n^{-\frac{3}{2}}) = \frac{1}{2\sigma_X^2}\frac{u^2}{n} + O(n^{-\frac{3}{2}}).$$

Note that $\frac{u}{\sqrt{n}}$ is a realization of a random variable $\frac{U}{\sqrt{n}} = M_n - p$, while s is a realization of M_n, where we treat M_n as a continuous random variable since its support given by $\{\frac{k}{n} | k = 0, 1, ..., n\}$ is a dense subset of $[0,1]$ in the limit. Further, observe that

$$nI_p(s) = \frac{u^2}{2\sigma_X^2} + O(n^{-\frac{1}{2}}) \to \frac{u^2}{2\sigma_X^2} = \frac{(u - E(U))^2}{2\sigma_U^2}, \tag{1.2}$$

where $E(U) = 0$ follows by definition of $\frac{U}{\sqrt{n}}$, and $\sigma_u^2 = \sigma_X^2$ since $Var(p + \frac{U}{\sqrt{n}}) = \frac{\sigma_U^2}{n} = Var(M_n) == \frac{\sigma_X^2}{n}$. Also, observe that if we take a deviation of order greater than $n^{-\frac{1}{2}}$ (e.g. if we replace $\frac{u}{\sqrt{n}}$ with $\frac{u}{n^{\frac{1}{4}}}$), the asymptotics in (1.2) will not hold, as we have claimed at the beginning. Moreover, we will see in Section 2.5 that the arguments

above readily apply to any iid sample mean. This suggests that the LDP is in line with the CLT.

Lastly, in Section 4 we will discuss the relationship between the LDP and LLN with more formality. Namely, we will prove that any random variable Y_n that satisfies the LDP with a twice differentiable strictly convex rate function $I(y)$ converges weakly in probability to an arbitrarily small open interval around $I^{-1}(0)$. From there we will obtain the Weak LLN for an iid sample mean, which will be readily applicable to our introductory example.

2 The Large Deviation Principle

Although the most general and rigorous formulation of the LDP involves a general topological space, in this paper we will limit ourselves to random variables defined on the real line. This limitation, however, is inconsequential for introductory purposes, since the essence of the LDP is invariant of the space it is defined on (we will see this in Section 2.4, where we give a precise statement (but not a rigorous formulation) of the LDP for a general topological space).

Before we turn to the meaning of the LDP, we devote a brief section to the idea of domination, which will allow us to describe the LDP intuitively without losing precision. Moreover, we will rely on the idea of domination to derive a formulation of the LDP in terms of tails of a probability distribution in Section 2.3, and to prove some results regarding the relationship between the pdf/pmf (probability density/mass function) of a random variable and the LDP in Section 3.

2.1 Domination

Let $f(n)$ and $g(n)$ denote two arbitrary functions of n. If $g(n)$ becomes much larger than $f(n)$ as $n \to \infty$, we say that $g(n)$ dominates $f(n)$ and write $f(n) = o(g(n))$, which reads as "$f(n)$ is little-o $g(n)$", or "$f(n)$ is of order less than $g(n)$". More precisely, this means that $lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$, which we formalize below using a trivial rearrangement of the precise definition of the limit.

Definition 2.1. ***(Domination)*** $f(n) = o(g(n))$ *iff* $\forall \varepsilon > 0\ \exists m_\varepsilon \in \mathbb{N}$ such that $|f(n)| < \varepsilon |g(n)|$ for $n > m_\varepsilon$.

This formulation is convenient, because intuitively it reads as: "eventually $|f(n)|$ is less than an arbitrarily small fraction of $|g(n)|$".

For an arbitrary expression $h_1(n)$, we say that $h_2(n)$ is a good asymptotic approximation of $h_1(n)$ and write $h_1(n) \approx h_2(n)$ iff in the limit the error of approximation becomes an arbitrarily small fraction of the approximated expression $h_1(n)$. We formalize this definition below.

Definition 2.2. ***(Good Asymptotic Approximation)*** *$h_1(n) \approx h_2(n)$ iff $h_1(n) - h_2(n) = o(h_1(n))$.*

We prove some elementary consequences of Definition 2.2 next.

Proposition 2.1. *The following are equivalent:*

(a) $h_1(n) - h_2(n) = o(h_1(n))$

(b) $\lim_{n\to\infty} \frac{h_2(n)}{h_1(n)} = 1$

(c) $h_1(n) - h_2(n) = o(h_2(n))$

Proof. Denote $h_{1,2}(n)$ by $h_{1,2}$. We prove the following implication chain: $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)$:

$(a) \Rightarrow (b) : \lim_{n\to\infty} \frac{h_2}{h_1} = \lim_{n\to\infty} \frac{h_1+(h_2-h_1)}{h_1} = \lim_{n\to\infty} \frac{h_1}{h_1} + \lim_{n\to\infty} \frac{(h_2-h_1)}{h_1} = 1+0 = 1.$

$(b) \Rightarrow (c) : \lim_{n\to\infty} \frac{h_1-h_2}{h_2} = \lim_{n\to\infty} \frac{\frac{h_1-h_2}{h_1}}{\frac{h_2}{h_1}} = \frac{\lim_{n\to\infty} \frac{h_1}{h_1} - \lim_{n\to\infty} \frac{h_2}{h_1}}{\lim_{n\to\infty} \frac{h_2}{h_1}} = \frac{1-1}{1} = 0.$

$(c) \Rightarrow (a) : \lim_{n\to\infty} \frac{h_1-h_2}{h_1} = \lim_{n\to\infty} \frac{\frac{h_1-h_2}{h_2}}{\frac{h_1+h_2-h_2}{h_2}} = \frac{\lim_{n\to\infty} \frac{h_1-h_2}{h_2}}{\lim_{n\to\infty} \frac{h_1-h_2}{h_2} + \lim_{n\to\infty} \frac{h_2}{h_2}} = \frac{0}{0+1} = 0.$

□

The meaning of the approximation in our introductory example should now be clearer. We prove a statement that will be used throughout the paper next.

Fact 2.1. *Observe that if $o_i(n)$, $i = 1, 2, .., m$, are m arbitrary expressions of order less than n, then $\lim_{n\to\infty} \frac{o_1(n)+...+o_m(n)}{n} = 0$. Thus, $\sum_{i=1}^{m} o_i(n) = o(n)$. We will use this fact to prove the following proposition.*

Proposition 2.2. *Let $g_1(n)$ and $g_2(n)$ be two arbitrary expressions of order less than n and let $0 < k_1 < k_2$, $k_{1,2} \in \mathbb{R}$. Then, $e^{-nk_2+g_2(n)} = o(e^{-nk_1+g_1(n)})$.*

Proof. Observe that $\frac{e^{-nk_2+g_2(n)}}{e^{-nk_1+g_1(n)}} = e^{-n(k_2-k_1)+g(n)}$, where $g(n) = g_2(n) - g_1(n) = o(n)$, by assumption and Fact 2.1. Thus, if we let $0 < \delta < k_2 - k_1$, we have by Definition 2.1

that $\exists N \in \mathbb{N}$ such that for $n > N$ $|g(n)| < \delta n$, which implies that

$$\begin{aligned} &g(n) < \delta n, \\ &g(n) - n(k_2 - k_1) < \delta n - n(k_2 - k_1) \\ &g(n) - n(k_2 - k_1) < -n(k_2 - k_1 - \delta). \end{aligned} \tag{2.1}$$

Since $\delta < k_2 - k_1$, $-n(k_2 - k_1 - \delta) \to -\infty$ as $n \to \infty$, and thus $g(n) - n(k_2 - k_1) \to -\infty$. It follows that $e^{-n(k_2-k_1)+g(n)} \to 0$ and $e^{-nk_2+g_2(n)} = o(e^{-nk_1+g_1(n)})$, as desired. □

2.2 Large Deviation Principle and Exponential Decay

In this section we give a narrow formulation of the LDP and relate it to the notion of exponential decay. As we will see in Section 2.4, the narrow formulation that we give below is only a sufficient condition for the LDP to hold. This section expands on the discussion of [6] regarding the little-o notation and the LDP.

Definition 2.3. *(**Narrow LDP**) A random variable Y_n satisfies the large deviation principle with rate function $I(y)$, if for any set $B \subset \mathbb{R}$ such that* $\Pr(Y_n \in B) \neq 0$

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n \in B)}{n} = -\inf{}_{y\in B} I(y).$$

We call $-\inf_{y\in B} I(y)$ the probability decay rate, or just a decay rate, of a set B. To relate Definition 2.3 to exponential decay, we prove the following fact:

Proposition 2.3. $\lim_{n\to\infty} \frac{\log P(Y_n\in B)}{n} = -\inf_{y\in B} I(y) \;\Leftrightarrow\; \Pr(Y_n \in B) = e^{-n(\inf_{y\in B} I(y))+o(n)}$.

Proof. Let $\inf_{y\in B} I(y) = I_B$. Assume $\Pr(Y_n \in B) = e^{-nI_B+o(n)}$. Then

$$\begin{aligned} &\lim{}_{n\to\infty} \frac{\log \Pr(Y_n \in B)}{n} = \lim_{n\to\infty} \frac{\log(e^{-nI_B+o(n)})}{n} = \\ =\ &\lim{}_{n\to\infty} \frac{-nI_B + o(n)}{n} = -I_B + \lim{}_{n\to\infty} \frac{o(n)}{n} = -I_B. \end{aligned}$$

Now assume $\lim_{n\to\infty} \frac{\log \Pr(Y_n\in B)}{n} = -I_B$. Since $\Pr(Y_n \in B) > 0$, we can represent it as

$$\Pr(Y_n \in B) = e^{-nI_B+(nI_B+\log \Pr(Y_n\in B))}.$$

Observe that

$$\frac{nI_B + \log \Pr(Y_n \in B)}{n} = I_B + \frac{\log \Pr(Y_n \in B)}{n},$$

where $\frac{\log \Pr(Y_n \in B)}{n} \to -I_B$, by assumption. Hence $\frac{nI_B + \log \Pr(Y_n \in B)}{n} \to 0$ and thus $nI_B + \log \Pr(Y_n \in B) = o(n)$, as desired. □

We write $\Pr(Y_n \in B) \asymp e^{-n(\inf_{y \in B} I(y))}$ to indicate that Y_n satisfies the LDP with rate function $I(y)$, where by $a_n \asymp b_n$ we mean that $\lim_{n\to\infty} \frac{\log a_n}{n} = \lim_{n\to\infty} \frac{\log b_n}{n}$. Notice that " $\asymp$ " is very different from " $\approx$ " discussed in Section 2.1.

Notice that if a random variable Y_n satisfies the LDP, it has to be that for any $B \subset \mathbb{R}$ $\inf_{y \in B} I(y) \geq 0$, for otherwise $\Pr(Y_n \in B) = e^{-n(\inf_{y \in B} I(y)) + o(n)}$ will eventually exceed 1. Therefore $I(y)$ is always non-negative. In general, it is possible for $I(y)$ to have more than one zero (see Section III.E.7 of [6], for instance). However, usually $I(y) = 0$ for a unique $y_0 \in \mathbb{R}$ and that is what we assume in this paper unless otherwise stated.

If $I(y)$ is continuous in the sense that $\lim_{y \to a} I(y) = I(a)$ and has a unique zero y_0, then by Proposition 2.3 it is clear that Definition 2.3 states that probability of any set that does not contain some open ball around y_0 decays exponentially fast at the rate of its (asymptotically) most likely elementary event (by elementary events we mean events of the form $\{Y_n = y_n\}$ in the descrete case and events of the form $\{Y_n \in (y, y + dy)\}$ in the continuous case). It then follows that the probability of an arbitrary open ball around y_0 is going to 1, i.e. probability gets concentrated around a single value in the support of Y_n. This suggests that Y_n satisfies the weak law of large numbers, which we discuss in more detail in Section 4.

Also notice that if $B_\varepsilon(y_0)$, $\varepsilon > 0$ denotes an arbitrary open ball around y_0, then the decay rate of $B_\varepsilon(y_0)$ is 0. This fact is natural and necessary since $\Pr(B_\varepsilon(y_0)) \to 1$. However, the convergence of $\Pr(B_\varepsilon(y_0))$ to1 is not a consequence of its decay rate being 0. If we look at the pmf of our introductory example, we see that $\Pr(S_n = \lfloor np \rfloor) \to 0$ even though its decay rate equals 0. This is because the decay rate is in essence *exponential*: it being zero simply means that *exponential* decay does not take place - it does not provide information about the direction of subexponential changes. This property can be inferred from Proposition 2.3 by observing that probability of a set with a decay rate of 0 is simply $e^{o(n)}$.

2.3 LDP In Terms of Strictly Convex Rate Functions and Tails of Probability Distribution

In some texts (e.g. [3]), the LDP is informally stated using the tails of probability distributions. In this section *we are going to give an original proof* that the

narrow LDP of Definition 2.3 can be formulated in terms of probability tails and strictly convex rate functions (such rate functions are very common as we will see in Section 2.5). Note that, as will be shown below, such a formulation is in line with Definition 2.3, but is not its equivalent, since not all rate functions are convex.

Lemma 2.1. *Let $\mathcal{S}_{Y_n}$ denote the support of a random variable Y_n and let $\mathcal{S}_Y = lim_{n\to\infty}\mathcal{S}_{Y_n}$. For any set $B \subset \mathbb{R}$, let $B_1 = B \cap (-\infty, y_0]$ and $B_2 = (y_0, \infty)$ so that $B = B_1 \cup B_2$. If $I : \mathcal{S}_Y \to \{0\} \cup \mathbb{R}^+$ is twice differentiable strictly convex function with $I(y_0) = 0$, then y_0 is a unique zero of $I(y)$ and for any set $B \subset \mathcal{S}_Y$*

$$\inf_{y\in B} I(y) = \min\{I(\sup B_1), I(\inf B_2)\}.$$

Proof. Recall from calculus, that since $I(y)$ is strictly convex, $I''(y) > 0\ \forall y \in \mathcal{S}_Y$. Notice that since $I(y) \geq 0$ it has to be that $I'(y_0) = 0$ (otherwise $I(y)$ will be negative to the left or to the right of y_0). Since $I'(y)$ is strictly increasing everywhere in its domain, $I'(y) < 0$ for $y < y_0$ and $I'(y) > 0$ for $y > y_0$. Simply put, $I(y)$ is strictly increasing as y goes away from y_0. The following results trivially follow:

$$\begin{aligned}\inf_{y\in B_1} I(y) &= I(\sup B_1\})\\ \inf_{y\in B_2} I(y) &= I(\inf B_2\})\\ I^{-1}(0) &= \{y_0\}\end{aligned}$$

where $\inf_{y\in B} I(y) = \min\{I(\sup B_1), I(\inf B_2)\}$ readily follows from above. □

Remark 2.1. Note that if $B = \{y_1\} \cup (y_2, y_3)$, $y_0 < y_1 < y_2 < y_3$, then for continuous Y_n with a pdf $f_n(y)$, $\Pr(Y_n \in B) = \int_{y_1}^{y_1} f_n(y)dy + \int_{y_2}^{y_3} f_n(y)dy = \int_{y_2}^{y_3} f_n(y)dy = \Pr(Y_n \in B')$, where $B' = (y_2, y_3)$. On the other hand, $\inf_{y\in B} I(y) = I(y_1) \neq I(y_2) = \inf_{y\in B'} I(y)$ and thus by Proposition 2.3 $\Pr(Y_n \in B) = e^{-nI(y_1)+o(n)} \neq \Pr(Y_n \in B') = e^{-nI(y_2)+o(n)}$. We resolve this discrepancy by noting that with sets like B $\inf I(y)$ should be taken over the interior, since there exists an open ball around y_1 that does not contribute to $\Pr(Y_n \in B)$. ■

Lemma 2.2. *Let $I(y)$ be defined as in Lemma 2.1 and let $y \geq I^{-1}(0)$. Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n > y)}{n} = -I(y) \ \Rightarrow\ \lim_{n\to\infty} \frac{\log \Pr(Y_n \in (y, y+\varepsilon))}{n} = -I(y).$$

Proof. Observe that by assumption and Proposition 2.3,

$$\Pr(Y_n \in (y, y+\varepsilon) = \Pr(Y_n > y) - \Pr(Y_n > y+\varepsilon) = e^{-nI(y)+o(n)} - e^{-nI(y+\varepsilon)+o(n)}. \quad (2.2)$$

Also note that

$$\Pr(Y_n \in (y, y+\varepsilon) - e^{-nI(y+\varepsilon)+o(n)} = e^{-nI(y)+o(n)} - 2e^{-nI(y+\varepsilon)+o(n)} < \Pr(Y_n \in (y, y+\varepsilon) \quad (2.3)$$

and

$$e^{-nI(y)+o(n)} - 2e^{-nI(y+\varepsilon)+o(n)} = e^{-nI(y)+o(n)} - o(e^{-nI(y)+o(n)}) \succ \frac{1}{2}e^{-nI(y)+o(n)}. \quad (2.4)$$

The equality in (2.4) follows from Proposition 2.2**,** and by " $\succ$ " we mean "is greater than for large enough n", which in turn follows from the meaning of little-o. Putting together (2.2), (2.3) and (2.4), we have that for large enough n,

$$\frac{1}{2}e^{-nI(y)+o(n)} < \Pr(Y_n \in (y, y+\varepsilon) < e^{-nI(y)+o(n)},$$

which, by Squeeze Theorem from calculus, implies that

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n \in (y, y+\varepsilon))}{n} = -I(y).$$

□

Remark 2.2. The proof above can easily be modified to show that for $y < I^{-1}(0)$ and $\varepsilon > 0$, $\lim_{n\to\infty} \frac{\log \Pr(Y_n < y)}{n} = -I(y) \Rightarrow \lim_{n\to\infty} \frac{\log \Pr(Y_n \in (y-\varepsilon, y))}{n} = -I(y)$.

Theorem 2.1. *Let $I(y)$ be a twice differentiable non-negative strictly convex function with a unique zero y_0. Then, a continuous random variable Y_n satisfies the LDP with rate function $I(y)$ if for any $y_1 \leq y_0$ and any $y_2 > y_0$*

$$lim_{n\to\infty} \frac{\log \Pr(Y_n < y_1)}{n} = -I(y_1)$$
$$lim_{n\to\infty} \frac{\log \Pr(Y_n > y_2)}{n} = -I(y_2)$$

Proof. Let $B = B_1 \cup B_2$ be an arbitrary set in the support of Y_n, where B_1 and B_2 are defined as in Lemma 2.1. Recall that, by Lemma 2.1, $\inf_{y\in B} I(y) = \min\{I(\sup B_1), I(\inf B_2)\}$. Without loss of generality, assume $I(\sup B_1) > I(\inf B_2)$, i.e. $\inf_{y\in B} I(y) = I(\inf B_2)$ (the

proof is identical for the other case). Moreover, assume $\inf B_2$ is not isolated, i.e. $\exists \varepsilon > 0$ such that $(\inf B_2, \inf B_2 + \varepsilon) \subset B_2$ (if $\inf B_2$ is isolated, consider instead $\inf int(B_2)$ as justified in Remark 2.1 and continue). Notice the following inequalities:

$$\Pr(Y_n \in [\inf B_2, \inf B_2 + \varepsilon) \leq \Pr(Y_n \in B_2) \leq \Pr(Y_n \in [\inf B_2, \infty).$$

By assumption and Lemma 2.2,

$$\lim_{n\to\infty} \frac{\log \Pr[Y_n \in [\inf B_2, \infty)]}{n} = -I(\inf B_2) = \lim_{n\to\infty} \frac{\log \Pr[Y_n \in [\inf B_2, \inf B_2 + \varepsilon)]}{n}.$$

Thus, we have by squeeze theorem from calculus that

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n \in B_2)}{n} = -I(\inf B_2).$$

Applying similar reasoning to B_1, we have that

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n \in B_1)}{n} = -I(supB_1).$$

Observe that B_1 and B_2 are disjoint and thus $P(B) = P(B_1) + P(B_2) = e^{-nI(\sup B_1)+o(n)} + e^{-nI(\inf B_2)+o(n)}$, where the last equality follows from Proposition 2.3. Since, by assumption, $I(\sup B_1) > I(\inf B_2)$, we have by Proposition 2.2 and the definition of little-o that for large enough n,

$$e^{-nI(\inf B_2)+o(n)} < P(B) < 2e^{-nI(\inf B_2)+o(n)}$$

which, by squeeze theorem from calculus, implies that

$$\lim_{n\to\infty} \frac{\log \Pr(B)}{n} = -I(\inf B_2) = -\inf{}_{y\in B} I(y), \tag{2.5}$$

where the last equality follows by our assumption at the beginning of the proof. Since B is an arbitrary set in the support of Y_n, (2.5) is exactly the statement of Definition 2.3 and the proof is complete. □

Remark 2.3. Note that the proof of Theorem 2.1 readily applies to discrete random variables if $I(y)$ can be obtained as twice differentiable function as in our introductory example. Moreover, whenever the support of Y_n can be obtained as a continuous one, one need not worry about the isolation of $\inf B_2$ as long as it is in the support of Y_n

(recall that we have obtained $[0,1]$ as a support of our discrete introductory example). ■

Recall that if Y_n is a continuous random variable with a pdf (probability density function) $f_n(y)$, then $\Pr(Y_n \in (y, y+dy)) = f(y)dy$. Thus, if Y_n satisfies the LDP with strictly convex rate function $I(y)$ it has to be by Proposition 2.3 and Lemma 2.1 that for $y > I^{-1}(0)$

$$f_n(y)dy = e^{-nI(y)+o(n)},$$

which implies that

$$f_n(y) = e^{-nI(y)+o(n)-\log(dy)} = e^{-nI(y)+o(n)}.$$

On the converse side, we will show in Section 3.2 that a continuous random variable Y_n satisfies the LDP with rate function $I(y)$, if its pdf admits the form $e^{-nI(y)+o(n)}$, where $I(y)$ is a non-negative twice differentiable strictly convex function and the expression under $o(n)$ does not depend on y, or is bounded in y. Moreover, we will see in Section 3.1 that a similar statement also applies to a pmf of a discrete random variable that satisfies the LDP.

2.4 Precise Statement of the Large Deviation Principle

As we have stated at the beginning, the formulation of the large deviation principle is essentially an attempt to abstract a particular property that many random variables exhibit. As we saw in the previous subsection, the LDP is a statement about exponential decay of asymptotically atypical events. The statement that we gave in Definition 2.3, however, is relatively narrow and does not encompass other common cases in which atypical events decay exponentially fast. A broader formulation of the LDP follows next. The *Large Deviation Principle* statement is taken from [5].

Definition 2.4. ***(Large Deviation Principle)*** *A random variable Y_n satisfies the large deviation principle with speed s_n and rate function $I(y)$ if and only if for any set $A \subset \chi$ the following two conditions are satisfied:*

$$\limsup_{n\to\infty} \frac{logP(Y_n \in A)}{s_n} \leq -\inf_{y\in\bar{A}} I(y)$$
$$\liminf_{n\to\infty} \frac{logP(Y_n \in A)}{s_n} \geq -\inf_{y\in int(A)} I(y).$$

χ denotes a general topological space. s_n is some expression such that $s_n \to \infty$ as $n \to \infty$. For a sequence a_n, $\liminf_{n\to\infty} a_n = \lim_{n\to\infty} inf\{a_k\}_{k\geq n}$ and $\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup\{a_k\}_{k\geq n}$. Let $i_n = \inf\{\frac{logP(Y_n\in A)}{s_n}\}_{k\geq n}$, $c_n = \sup\{\frac{logP(Y_n\in A)}{s_n}\}_{k\geq n}$. Also let $L_1 = \liminf_{n\to\infty} \frac{logP(Y_n\in A)}{s_n}$ and $L_2 = \limsup_{n\to\infty} \frac{logP(Y_n\in A)}{s_n}$.

Observe that $i_n = L_1 + i_n - L_1$, where $i_n - L_1 \to 0$, since $i_n \to L_1$. Thus, $i_n = L_1 + o(1)$. Likewise, $c_n = L_2 + o(1)$. Hence, $\forall n \in \mathbb{N}$

$$L_1 + o(1) \leq \frac{\log \Pr(Y_n \in A)}{s_n} \leq L_2 + o(1).$$

Observe that $o(1) \times s_n = o(s_n)$, for $\frac{o(1)\times s_n}{s_n} = o(1) \to 0$. Thus, the inequalities above are equivalent to

$$e^{s_n L_1 + o(s_n)} \leq \Pr(Y_n \in A) \leq e^{s_n L_2 + o(s_n)}.$$

Hence, if we let $I_{A+} = \inf_{y\in\bar{A}} I(y)$ and $I_{A-} = \inf_{y\in int(A)} I(y)$, we have that

$$e^{-s_n I_{A-} + o(s_n)} \leq \Pr(Y_n \in A) \leq e^{-s_n I_{A+} + o(s_n)}.$$

Observe that if $s_n = n$ and $I_{A-} = I_{A+}$, then $\inf_{y\in\bar{A}} I(y) = \inf_{y\in int(A)} I(y) = \inf_{y\in A} I(y)$ and thus

$$\Pr(Y_n \in A) = e^{-n(\inf_{y\in A} I(y)) + o(n)}. \tag{2.6}$$

Hence, the LDP as formulated in Definition 2.3 is just a special case of a broader formulation that we have above. Also note that (2.6) can hold even when $I_{A-} \neq I_{A+}$ if $L_1 = L_2 = L$, in which case $\lim_{n\to\infty} \frac{\log \Pr(Y_n\in A)}{n} = L$. For instance, if we consider our Bernoulli iid sample mean M_n from Section 1 in light of sets of the form $A = \{\frac{\lfloor ns_1 \rfloor}{n}\} \cup (\frac{\lfloor ns_2 \rfloor}{n}, \frac{\lfloor ns_3 \rfloor}{n})$, $p < s_1 < s_2 < s_3$, then we have that $\inf_{s\in int(A)} I(s) = I(s_2) \neq \inf_{s\in\bar{A}} I(s) = I(s_1)$. Even so, we will see in Section 3.1 that $\lim_{n\to\infty} \frac{\log \Pr(M_n\in A)}{n} = -\inf_{s\in A} I(s) = -I(s_1)$.

2.5 Gärtner-Ellis Theorem and iid Sample Mean

Two of the most important goals of the Large Deviations Theory is identifying cases in which the LDP is satisfied and coming up with ways of calculating rate functions (combinatorial approaches are not always effective or efficient). Gärtner-Ellis (GE) Theorem is one of the most important results in this direction, for it accomplishes both tasks for a large class of random variables.

Let Y_n be a random variable defined on $\mathbb{R}^d$ and let the scaled cumulant generating function of Y_n be defined as $\lambda(k) = \lim_{n\to\infty} \frac{logE(e^{knY_n})}{n}$. We give the statement of the Gärtner-Ellis Theorem as formulated in [7] below.

Theorem 2.2. *If $\lambda(k)$ is finite, and differentiable for all $k \in \mathbb{R}^d$, then Y_n satisfies the LDP with rate function*

$$I(y) = \sup_{k\in\mathbb{R}^d} \{ky - \lambda(k)\}.$$

Note that the rate function of our introductory example can alternatively be obtained through GE theorem: using iid properties one can easily show that in case of an iid sample mean $\lambda(k)$ is reduced to the log of moment-generating function of the underlying distribution; the rate function can then be easily obtained using methods from calculus. See [6] for examples and a detailed discussion of the Gärtner-Ellis Theorem. Here we simply state basic facts that follow from GE theorem above:

- Any rate function obtained through GE is strictly convex.
- Any iid sample mean with underlying random variable X satisfies the LDP with speed $s_n = n$ and rate function $I(x)$. Further, letting $E(X) = \mu$, we have that

$$I(\mu) = 0,\ I'(\mu) = 0,\ I''(\mu) = \frac{1}{\sigma_X^2}.$$

The second fact shows that the comment about the LDP and CLT made in Section 1 applies to any iid sample mean.

3 Inferring Large Deviations From Density/Mass Functions

By now it should be clear that if a discrete random variable Y_n satisfies the LDP, its pmf has to be a decaying exponential in n for atypical values of Y_n. We also said in Section 2.3 that if $f_n(y)$ is a pdf of a continuous random variable Y_n that satisfies the LDP, then $f_n(y)dy$ - the probability of an infinitesimal interval not containing a typical value of Y_n - decays exponentially fast. Note, however, that these are only neccessary conditions for the LDP to hold, for by Definition 2.3, probability of any set not containing a typical

value of Y_n has to be decaying exponentially fast. This, however, is not immediate from above, for it is not obvious that the sum of infinitely many decaying exponentials decays exponentially fast. In the two subsections that follow we address the discrete and continuous cases separately and prove sufficient conditions under which the infinite sum of decaying exponentials indeed decays exponentially fast. *Proofs that follow in the two subsections are original.*

3.1 The Discrete Case

Recall from our introductory example that if S_n represents the sum of n iid Bernoulli trials each with success probability p, then any event of the form $\{\frac{S_n}{n} = \frac{\lfloor ns \rfloor}{n}\}$, $s \neq p$, decays exponentially fast. However, it does not readily follow that $\frac{S_n}{n}$ satisfies the LDP as formulated in Definition 2.3, for one needs to show that the tails of a probability distribution of $\frac{S_n}{n}$ decay exponentially fast. This fact is not immediate, because the support of $\{\frac{S_n}{n} > \frac{\lfloor ns \rfloor}{n}\}$ given by $\{\frac{\lfloor ns \rfloor + 1}{n}, \frac{\lfloor ns \rfloor + 2}{n}, ..., \frac{\lfloor ns \rfloor + n - \lfloor ns \rfloor}{n}\}$ contains infinitely many terms in the limit, for the number of elements in the support $n - \lfloor ns \rfloor \to \infty$. In fact, after we expose the proof that our introductory example satisfies the LDP, we will show that some infinite sums of decaying exponentials do not decay at all, and will see how an infinum of a rate function affects the asymptotic behavior of a probability of a set. Lastly, we will generalize our proof to derive sufficient conditions for a discrete random variable to satisfy the LDP.

Theorem 3.1. *Let S_n be the sum of n iid Bernoulli trials each with success probability p and let $I(s)$ be the rate function obtained in Section 1. Then, for $s \in (p, 1)$*

$$\lim_{n \to \infty} \frac{\log \Pr(S_n > \lfloor ns \rfloor)}{n} = -I(s).$$

Proof. Observe that for $s \in [0,1], \lfloor ns \rfloor \in \{0, 1, 2, ..., n\}$ for any given n. Also recall that for $s = 0, 1$ one need not apply the approximation in 1.1, for it is immediate that $P(S_n = \lfloor ns \rfloor) = \binom{n}{\lfloor ns \rfloor} p^{\lfloor ns \rfloor}(1-p)^{n - \lfloor ns \rfloor} = e^{-nI(s)}$ (in fact, the approximation is not defined for $s = 0, 1$). Thus, for any n the values of $\lfloor ns \rfloor$ to which approximation is applied range from 1 to $n-1$. It can be shown using this fact that $\log(\alpha_n \beta_n)$ is bounded in s by a sublinear expression in n (recall from Section 1 that $P(S_n = \lfloor ns \rfloor) \approx e^{-nI_p(s) + \log(\alpha_n \beta_n)} = e^{-nI_p(s) + o_s(n)}$, where $o_s(n)$ denotes a sublinear expression in n that also depends on s).

Let $\bar{o}(n)$ denote an upper bound of $\log(\alpha_n\beta_n)$ in s that is of order less than n. Further, observe that

$$P(\frac{S_n}{n} = \frac{\lfloor ns \rfloor}{n}) = P(S_n = \lfloor ns \rfloor) \approx e^{-nI_p(s)+o_s(n)} \leq e^{-nI_p(s)+\bar{o}(n)}.$$

Also note that the support of $\{S_n > \lfloor ns \rfloor\}$ is given by $\{\lfloor ns \rfloor + 1, \lfloor ns \rfloor + 2, ..., \lfloor ns \rfloor + n - \lfloor ns \rfloor\}$. We then have that for $s \in (p, 1)$

$$P(S_n > \lfloor ns \rfloor) = \sum_{i=1}^{n-\lfloor ns \rfloor} P(S_n = \lfloor \left(\frac{\lfloor ns \rfloor + i}{n} \right) n \rfloor) \approx \sum_{i=1}^{n-\lfloor ns \rfloor} e^{-nI(\frac{\lfloor ns \rfloor + i}{n})+o_i(n)}.$$

Recall from Lemma 2.1 that since $s > p$, I attains its minimum at $\frac{\lfloor ns \rfloor + 1}{n}$ for any given n. Thus, if $o_1(n)$ denotes a sublinear term in the expression for $P(S_n = \lfloor \left(\frac{\lfloor ns \rfloor + 1}{n} \right) n \rfloor)$, we have that

$$e^{-nI(\frac{\lfloor ns \rfloor + 1}{n})+o_1(n)} < P(S_n > \lfloor ns \rfloor) < ne^{-nI(\frac{\lfloor ns \rfloor + 1}{n})+\bar{o}(n)} = e^{-nI(\frac{\lfloor ns \rfloor + 1}{n})+o(n)},$$

where $o(n)$ in the rightmost expression is a function of n only. It is then immediate from inequalities above that

$$-I(\frac{\lfloor ns \rfloor + 1}{n}) + \frac{o_1(n)}{n} < \frac{\log \Pr(S_n > \lfloor ns \rfloor)}{n} < -I(\frac{\lfloor ns \rfloor + 1}{n}) + \frac{o(n)}{n}. \tag{3.1}$$

Further, note that $\frac{ns}{n} < \frac{\lfloor ns \rfloor + 1}{n} \leq \frac{ns+1}{n}$, where both bounds equal s in the limit. Thus, as $n \to \infty$, $\frac{\lfloor ns \rfloor + 1}{n} \to s$ and, by continuity of I, $I(\frac{\lfloor ns \rfloor + 1}{n}) \to I(s)$ (recall from Section 1 that I is differentiable). Thus, by definition of $o(n)$, both bounds in (3.1) go to $-I(s)$ and we have by squeeze theorem from calculus that

$$\lim_{n \to \infty} \frac{\log \Pr(S_n > \lfloor ns \rfloor)}{n} = -I(s),$$

as desired. □

Corollary 3.1. *$\frac{S_n}{n}$ satisfies the LDP with rate function $I(s)$.*

Proof. Observe that events $\{S_n > \lfloor ns \rfloor\}$ and $\{S_n > ns\}$ have the same support and thus $\Pr(S_n > \lfloor ns \rfloor) = \Pr(S_n > ns) = \Pr(M_n > s)$, where $M_n = \frac{S_n}{n}$. Hence, $\lim_{n\to\infty} \frac{\log \Pr(M_n > s)}{n} = -I(s)$. Moreover, modifing the proof of Theorem 3.1 for $\Pr(S_n < ns)$, $s \in (0, p)$, one can easily show that $\lim_{n\to\infty} \frac{\log \Pr(M_n < s)}{n} = -I(s)$. It then follows from Theorem 2.1 and

Remark 2.3 that M_n satisfies the LDP with rate function $I(s)$. □

Remark 3.1. Note that if we modify the proof of Theorem 3.1 for $s \in (p-\varepsilon, p+\varepsilon)$, $\varepsilon > 0$, we will get that $\lim_{n\to\infty} \frac{\log \Pr(M_n \in (p-\varepsilon,p+\varepsilon))}{n} = -I(p) = 0$, which yields no information about the behavior of $\Pr(M_n \in (p-\varepsilon, p+\varepsilon)$, as we have explained in Section 2.2. It then follows from Corrolary 3.1 that $\Pr(M_n \in (p-\varepsilon, p+\varepsilon) \to 1$ and so does $\Pr(M_n \in (p-\varepsilon, p) \cup (p, p+\varepsilon))$, since $\Pr(S_n = \lfloor np \rfloor) \to 0$. This fact is interesting, for like $\Pr(M_n > s)$, $s \in (p,1)$, $\Pr(M_n \in (p-\varepsilon, p) \cup (p, p+\varepsilon))$ is the sum of infinitely many decaying exponentials in the limit. The difference between the asymptotic behavior of the two can be attributed to the fact that arbitrarily small decay rates of exponentials that appear in the expression for $\Pr(M_n \in (p-\varepsilon, p) \cup (p, p+\varepsilon))$ as $n \to \infty$ fail to offset the effect of the summation of the enlarging series. ■

Lastly, we state a corrolary that readily follows from the discussions in this section.

Corollary 3.2. *A discrete random variable Y_n with support $\mathcal{S}_{Y_n}$ and pmf $p_n(y)$ satisfies the LDP if the following four conditions are satisfied:*

(1) $p_n(y) = e^{-nI(y)+o(n)}$, where $I(y)$ is a non-negative twice differentiable strictly convex function with $I^{-1}(0) \neq \emptyset$;

(2) $o(n)$ term in the expression of $p_n(y)$ does not depend on y or is bounded in y;

(3) $|\mathcal{S}_{Y_n}| = e^{o(n)}$;

(4) $\lim_{n\to\infty} y_{1,2}(n) = y$, where $y_1(n) = \min\{(y,\infty) \cap \mathcal{S}_{Y_n}\}$ and $y_2(n) = \max\{(-\infty, y) \cap \mathcal{S}_{Y_n}\}$.

3.2 The Continuous Case

We claimed in Section 2.3 that if a continuous random variable Y_n satisfies the LDP with strictly convex rate function $I(y)$, then its pdf can be written as $e^{-nI(y)+o(n)}$. In this section, we are going to prove that any continuous random variable Y_n satisfies the LDP with rate function $I(y)$, if its pdf admits the form $e^{-nI(y)+o(n)}$, where $I(y)$ is strictly convex. We first prove the case in which the expression under $o(n)$ does not depend on y, and then extend the result to the case in which $o(n)$ term depends on but is bounded in y. In the proofs that follow we assume that the support of Y_n is $\mathbb{R}$.

Lemma 3.1. *Let $I(y)$ be a twice differentiable strictly convex function on $[c,\infty)$, $c > I^{-1}(0)$. Let $\triangle_k I = I(c+k\varepsilon) - I(c+k\varepsilon - \frac{\varepsilon}{2})$, $\varepsilon > 0$, $k \in \mathbb{N}$. Then, $\triangle_{k+1} I > \triangle_k I$.*

Proof. Notice that for $a \geq c$, $I(a+\varepsilon) = I(a) + \int_a^{a+\varepsilon} I'(t)dt = I(a) + I(a+\varepsilon) - I(a)$. We know from calculus that a twice differentiable function is strictly convex on $[c,\infty)$ iff $I''(y) > 0$, $y \in [c,\infty)$. Thus, for $y > a \geq c$, $I'(a) < I'(y)$ and so $\varepsilon I'(a) = \int_a^{a+\varepsilon} I'(a)dt < \int_a^{a+\varepsilon} I'(t)dt$. Hence,

$$I(a+\varepsilon) > I(a) + \varepsilon I'(a)$$

and

$$I(a+\frac{\varepsilon}{2}) > I(a-\frac{\varepsilon}{2}) + \varepsilon I'(a-\frac{\varepsilon}{2})$$

Subtracting the two expressions, we get

$$I(a+\varepsilon) - I(a+\frac{\varepsilon}{2}) > I(a) - I(a-\frac{\varepsilon}{2}) + \varepsilon(I'(a) - I'(a-\frac{\varepsilon}{2})).$$

If $c < a - \frac{\varepsilon}{2}$, $I'(a) - I'(a-\frac{\varepsilon}{2}) > 0$ (by strict convexity of I) and thus $I(a+\varepsilon) - I(a+\frac{\varepsilon}{2}) > I(a) - I(a-\frac{\varepsilon}{2})$. Replacing a with $c + k\varepsilon$, we have that $c < c + k\varepsilon - \frac{\varepsilon}{2}$, $k \in \mathbb{N}$, and thus we get

$$I(c+(k+1)\varepsilon) - I(c+(k+1)\varepsilon - \frac{\varepsilon}{2}) > I(c+k\varepsilon) - I(c+k\varepsilon - \frac{\varepsilon}{2}),$$

which is exactly $\triangle_{k+1} I > \triangle_k I$. □

Corollary 3.3. *With $n \in \mathbb{N}$, $e^{-n\triangle_k I}$ is a decreasing sequence in k.*

Proof. This follows immediately from Lemma 3.1. □

Lemma 3.2. $A_{k+1} = o(A_k)$, $A_k = \int_{c+k\varepsilon}^{c+(k+1)\varepsilon} e^{-nI(t)+o(n)}dt$, $k \in \{0\} \cup \mathbb{N}$, $c > I^{-1}(0)$.

Proof. Since $I'(y) > 0$ for $y \in [c,\infty)$, we can denote an upper and a lower bound of A_k by $A_k^{UB} = \varepsilon e^{-nI(c+k\varepsilon)+o(n)}$ and $A_k^{LB} = \frac{\varepsilon}{2} e^{-nI(c+k\varepsilon+\frac{\varepsilon}{2})+o(n)}$, respectively. Thus, we have that

$$\lim_{n\to\infty} \frac{A_{k+1}^{UB}}{A_k^{LB}} = \lim_{n\to\infty} \frac{\varepsilon e^{-nI(c+(k+1)\varepsilon)+o(n)}}{\frac{\varepsilon}{2} e^{-nI(c+k\varepsilon+\frac{\varepsilon}{2})+o(n)}} = \tag{3.2}$$
$$= \lim_{n\to\infty} 2e^{-n\triangle_{k+1} I} = 0,$$

where the last equality follows from $\triangle_{k+1} I > 0$. Rewriting (3.2) using the precise statement of the limit, we get

$$\forall \varepsilon > 0 \quad \exists m_\varepsilon \in \mathbb{N} \ \textit{such that} \ A_{k+1}^{UB} < \varepsilon A_k^{LB} \quad \textit{for} \ n > m_\varepsilon.$$

Since $A_{k+1} < A_{k+1}^{UB}$ and $A_k^{LB} < A_k$, we have that

$$\forall \varepsilon > 0 \quad \exists m_\varepsilon \in \mathbb{N} \ \ such \ \ that \ \ A_{k+1} < \varepsilon A_k \quad for \ \ n > m_\varepsilon,$$

which is the desired result. □

Lemma 3.3. *For any* $\varepsilon > 0$, $\int_c^\infty e^{-nI(t)+o(n)}dt \approx \int_c^{c+\varepsilon} e^{-nI(t)+o(n)}dt$.[1]

Proof. Notice that $\int_c^\infty e^{-nI(t)+o(n)}dt = \sum_{k=0}^\infty \int_{c+k\varepsilon}^{c+(k+1)\varepsilon} e^{-nI(t)+o(n)}dt = A_0 + \sum_{k=1}^\infty A_k$. To prove this Lemma, we need to show that $\sum_{k=1}^\infty A_k = o(A_0)$. As we have shown in the proof of Lemma 3.2,

$$\exists m \in \mathbb{N} \ \ such \ \ that \ \ for \ \ n > m \ \ \frac{A_1^{UB}}{A_0^{LB}} = 2e^{-n\triangle_1 I} < \delta, \ \ 0 < \delta < 1.$$

Thus, by Corrolary 3.3,

$$\forall k \in \mathbb{N} \ \ \frac{A_{k+1}^{UB}}{A_k^{LB}} < \delta \ \ for \ \ n > m,$$

which implies that

$$\forall k \in \mathbb{N} \ \ A_{k+1} < \delta A_k \ \ for \ \ n > m.$$

It then follows that for $n > m$,

$$A_k < \delta A_{k-1} < \delta^2 A_{k-2} < ... < \delta^{k-1} A_1.$$

Therefore,

$$\sum_{k=1}^\infty A_k = A_1 + \sum_{k=2}^\infty A_k < A_1 + \sum_{k=2}^\infty \delta^{k-1} A_1 = A_1(1 + \sum_{k=2}^\infty \delta^{k-1}).$$

Notice that $(1 + \sum_{k=2}^\infty \delta^{k-1})$ is an infinite geometric series with $0 < \delta < 1$. Thus, $(1 + \sum_{k=2}^\infty \delta^{k-1}) = \frac{1}{1-\delta}$. Hence, we have that

$$0 < \sum_{k=1}^\infty A_k < A_1(\frac{1}{1-\delta}).$$

It then follows by Lemma 3.2, $\lim_{n\to\infty} \frac{A_1(\frac{1}{1-\delta})}{A_0} = 0$. Hence, by Squeeze Theorem from

[1] Recall from Section 2.1 that $h_1 \approx h_2$ means that h_2 is a good asymptotic approximation of h_1, i.e. $h_1 - h_2 = o(h_2)$.

calculus, $\frac{\sum_{k=1}^{\infty} A_k}{A_0} \to 0$ as $n \to \infty$ and thus $\sum_{k=1}^{\infty} A_k = o(A_0)$. □

Theorem 3.2. *Let $I(y)$ be a non-negative twice differentiable strictly convex function with $I^{-1}(0) = y_0$. If a pdf of a random variable Y_n admits the form $e^{-nI(y)+o(n)}$, where $o(n)$ does not depend on y, then for $c > y_0$*

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n > c)}{n} = -I(c).$$

Proof. By Lemma 3.3, for any $\varepsilon > 0$,

$$\Pr(Y_n > c) = \int_c^{\infty} e^{-nI(t)+o(n)} dt \approx \int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt,$$

or equivalently,

$$\Pr(Y_n > c) = \int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt + o(\int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt). \tag{3.3}$$

By definition of little-o, $\exists N \in \mathbb{N}$ such that for $n > N$

$$o(\int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt) < \int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt$$

and thus

$$\Pr(Y_n > c) < 2\int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt. \tag{3.4}$$

We have shown in the proof of Lemma 2.1, that $I(y)$ is strictly increasing on $(I^{-1}(0), \infty)$. Since $c > I^{-1}(0)$, we have the following inequalities:

$$\frac{\varepsilon}{n} e^{-nI(c+\frac{\varepsilon}{n})+o(n)} < \int_c^{c+\varepsilon} e^{-nI(t)+o(n)} dt < \varepsilon e^{-nI(c)+o(n)}. \tag{3.5}$$

It then follows from 3.3, 3.4 and 3.5 that for large enough n,

$$\frac{\varepsilon}{n} e^{-nI(c+\frac{\varepsilon}{n})+o(n)} < P(Y_n > c) < 2\varepsilon e^{-nI(c)+o(n)}.$$

Further, observe that

$$\frac{\log(\frac{\varepsilon}{n} e^{-nI(c+\frac{\varepsilon}{n})+o(n)})}{n} = \frac{\log\varepsilon - \log n - nI(c+\frac{\varepsilon}{n}) + o(n)}{n} =$$
$$= -I(c+\frac{\varepsilon}{n}) + \frac{o(n)}{n} \equiv L_1$$

and

$$\frac{\log(2\varepsilon e^{-nI(c)+o(n)})}{n} = \frac{\log(2\varepsilon) - nI(c) + o(n)}{n} =$$
$$= -I(c) + \frac{o(n)}{n} \equiv L_2.$$

Also notice that

$$L_1 < \frac{\log \Pr(Y_n > c)}{n} < L_2.$$

Recall from calculus that since $I(y)$ is differentiable, it is continuous, i.e. $\lim_{y\to c} I(y) = I(c)$. Thus, $\lim_{n\to\infty} L_1 = -I(c)$. Observe also that $lim_{n\to\infty} L_2 = -I(c)$. It then follows from Squeeze Theorem from calculus that

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n > c)}{n} = -I(c),$$

as desired. □

Corollary 3.4. *Let $f_n(y)$ be a pdf of a random variable Y_n and let $I(y)$ be defined as in Theorem 3.2. If $f_n(y) = e^{-nI(y)+o_y(n)}$, where $o_y(n)$ is a sublinear expression in n that depends on but is bounded in y, then for $c > I^{-1}(0)$*

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n > c)}{n} = -I(c).$$

Proof. Let $\bar{o}(n)$ and $\underline{o}(n)$ denote an upper and a lower bound of $o_y(n)$ in y, respectively. Then, we have that

$$e^{-nI(y)+\underline{o}(n)} \leq e^{-nI(y)+o_y(n)} \leq e^{-nI(y)+\bar{o}(n)}.$$

Following the reasoning in the proof of Theorem 3.1,

$$\lim_{n\to\infty} \frac{\log \int_c^\infty e^{-nI(t)+\underline{o}(n)} dt}{n} = \lim_{n\to\infty} \frac{\log \int_c^\infty e^{-nI(t)+\bar{o}(n)} dt}{n} = -I(c).$$

Hence, by squeeze theorem from calculus,

$$\lim_{n\to\infty} \frac{\log \int_c^\infty e^{-nI(t)+o_t(n)} dt}{n} = \lim_{n\to\infty} \frac{\log \Pr(Y_n > c)}{n} = -I(c).$$

□

Corollary 3.5. *Let $f_n(y)$ be the pdf of a random variable Y_n and let $I(y)$ be defined as in Theorem 3.2. If $f_n(y) = e^{-nI(y)+o(n)}$, where the $o(n)$ term does not depend on y or is bounded in y, then Y_n satisfies the LDP with rate function $I(y)$.*

Proof. Observe that all the results of this section also hold for a lower tail, since $I(y)$ behaves in exactly the same way as y moves away from $I^{-1}(0)$ to the left. Thus, we have that for any $y_1 \leq y_0$ and any $y_2 > y_0$

$$\lim_{n\to\infty} \frac{\log \Pr(Y_n < y_1)}{n} = -I(y_1)$$
$$\lim_{n\to\infty} \frac{\log \Pr(Y_n > y_2)}{n} = -I(y_2).$$

By Theorem 2.1, Y_n satisfies the LDP with rate function $I(y)$. □

Example 3.1. It is well-known that the pdf of an iid sample mean M_n with underlying normally distributed random variable X with $E(X) = \mu$ can be obtained in the form $f_n(x) = \sqrt{\frac{n}{2\pi\sigma_X^2}} e^{-n\frac{(x-\mu)^2}{2\sigma_X^2}}$, where $\log\sqrt{\frac{n}{2\pi\sigma_X^2}} = o(n)$ does not depend on x and $\frac{(x-\mu)^2}{2\sigma_X^2}$ is non-negative, twice differentiable strictly convex function. Hence, M_n satisfies the LDP with rate function $\frac{(x-\mu)^2}{2\sigma_X^2}$.

4 Weak Law of Large Numbers and the Large Deviation Principle

In this section we expand on and formalize the discussion of [3] about the relationship between the LDP and the Weak LLN. We first give a precise statement of the Weak LLN for an iid sample mean and then formulate its generalization. Afterwards, *we give an original proof* that in many cases the LDP implies a generalized Weak LLN and will use this result to obtain the proof of Weak LLN for an iid sample mean.

The statement that follows is taken from [1].

Fact 4.1. (Weak LLN) Let (X_n) be a sequence of iid random variables. Let $M_n = \frac{1}{n}(X_1 + .. + X_n)$ and let $\mu = E(X) = E(M_n)$. Then, $\forall \varepsilon > 0$

$$\lim_{n\to\infty} \Pr(|M_n - \mu| < \varepsilon) = 1,$$

or equivalently,

$$\lim_{n\to\infty} \Pr(|M_n - \mu| > \varepsilon) = 0.$$

Weak LLN formulated above can be generalized in the following way:

Definition 4.1. ***(Generalized Weak LLN)*** *Let $Y_n = F(X_1, ..., X_n)$ denote some statistic of n-sized sample. Then, Y_n satisfies the LLN if the support of Y_n contains a point y^* such that $\forall \varepsilon > 0$*

$$\lim_{n\to\infty} \Pr(|Y_n - y^*| < \varepsilon) = 1,$$

or equivalently,

$$\lim_{n\to\infty} \Pr(|Y_n - y^*| > \varepsilon) = 0.$$

We prove the following proposition next.

Proposition 4.1. *The generalized Weak LLN holds for any random variable Y_n that satisfies the LDP with a twice differentiable strictly convex rate function $I(y)$, with y^* given by $I^{-1}(0)$.*

Proof. Take any $\varepsilon > 0$. Recall from Section 2.4 that if Y_n satisfies the LDP with rate function $I(y)$ and speed s_n, then for any set A in the support of Y_n

$$e^{-s_n I_{A-} + o(s_n)} \leq \Pr(Y_n \in A) \leq e^{-s_n I_{A+} + o(s_n)}, \tag{4.1}$$

where $\lim_{n\to\infty} s_n = \infty$, $I_{A+} = \inf_{y\in\bar{A}} I(y)$ and $I_{A-} = \inf_{y\in int(A)} I(y)$. Let $A = (-\infty, y_0 - \varepsilon) \cup (y_0 + \varepsilon, \infty)$, where $I(y_0) = 0$ (recall from Section 2.2 that by definition $I(y)$ has at least one zero). We then have by Lemma 2.1 that y_0 is a unique zero and $0 < I_{A+} \leq I_{A-}$ (the latter relies on the observation that $y_0 \notin \bar{A}$). This implies that $-s_n I_{A-} + o(s_n) \to -\infty$ and $-s_n I_{A+} + o(s_n) \to -\infty$, since $s_n \to \infty$ (see the proof of Proposition 2.2 for details, whose argument can readily applied to this case by considering s_n in place of n). It then follows that both bounds in (4.1) shrink to 0 as $n \to \infty$ and thus $\Pr(Y_n \in A) \to 0$ by squeeze theorem from calculus. Observing that $A = \{y \,|\, |y - y_0| > \varepsilon\}$ and y_0 plays the role of y^* in Definition 4.1, the proof is complete. □

Corollary 4.1. *(Weak LLN for an iid Sample Mean)* Recall from Section 2.5 that an iid sample mean M_n with underlying random variable X satisfies the LDP with speed n and twice differentiable strictly convex rate function $I(x)$, where $I^{-1}(0) = E(X) = \mu$. By Proposition 4.1, Weak LLN for an iid sample mean follows immediately.

References

[1] R. V. Hogg; J. W. McKean; A. T. Craig;. *Introduction To Mathematical Statistics*. Pearson, 2013.

[2] F. Rezakhanlou. Lectures on the large deviation principle. Electronic, March 2017. Obtained from the website of University of California, Berkeley.

[3] J. T. Lewis; R. Russell;. An introduction to large deviations for teletraffic engineers. Electronic, November 1997. Obtained from the website of University of Warwick.

[4] F. Rassoul-Agha; T. Seppalainen;. *A Course on Large Deviations with an Introduction to Gibbs Measures*. American Mathematical Society, 2015.

[5] J. M. Swart. Large deviation theory. Electronic, January 2012. Obtained from the website of UTIA (Institute of Information Theory and Automation of the Czech Academic Sciences).

[6] H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478:1–69, 2009.

[7] G. Wainrib. *Stochastic Biomathematical Models With Applications to Neuronal Modeling*, chapter A Brief Introduction to Large Deviations Theory, pages 57–72. Springer, 2014.