An Introduction to Large Deviation Theory

Daniel P. Gause

Department of Mathematics, Middlebury College, Middlebury, VT, 05753

(Dated: May 12, 2020)

Abstract

While any self-proclaimed probabilist will be familiar with the central limit theorem and the law of large numbers, these results only apply to the central tendencies of distributions. Large deviation theory extends probability theory to the tails of distributions of sums of independent identically distributed random variables. We find that for large deviations from the expected value of a series of iid random variables, there exist exponential decay functions describing probabilistic behavior in the tails of these distributions. We prove that the rate function for a series of iid random variables is the Fenchel-Legendre transform of the logarithm of the distribution's moment-generating function. While large deviation theory applies to much more complex systems of random variables, we will restrict the scope of our study to series of iid random variables.

CONTENTS

I. INTRODUCTION AND BACKGROUND

Random variables are the heart of probability theory. These variables give form to probability theory, which in turn informs us on their underlying properties, giving order to their randomness. We develop parameters such as expectation values and probabilistic bounds to classify their distributions. Then, by observing series of these random variables, we can further bound their randomness. The stochastic nature of individual variables smooths out into familiar functions as more are observed in a system, often leading us to continuity and differentiability. The limit behavior of a series of random variables holds important and applicable information. Yet with all of this theory constraining predictions, what happens when the unpredicted occurs? What happens when the stochasticity of random variables exceeds that accounted for in our models?

Herein lies large deviation theory (LDT) – the study of the bound probability of a sum of random variables differing substantially from its expected value. Though traces of LDT

show up in earlier work, it was first directly referenced by Harald Cramér in its relevancy to the insurance industry, namely to ruin theory. Cramér used LDT in the 1930s to describe an insurer's vulnerability to insolvency, where the random variables of interest were insured accidents. Cramér's work and subsequent research on LDT was presented in a unified formalization published in 1966 by Varadhan. Large deviation theory offers additional information intrinsic in probabilistic models, and therefore is widely applicable. While familiar principles such as the central limit theorem and the law of large numbers define the central tendencies of distributions, LDT explores the tails of these distributions. Principles of LDT have been used in areas of finance, information systems, and physics. LDT is directly related to entropy in statistical mechanics, and can be used to describe Brownian Motion. The applications are *boundless*.

This paper will serve as an accessible introduction to large deviation theory. It will cover the necessary preliminary theory, standardizing the notation of probability theory and introducing relevant theorems and laws. It will examine the fundamental concept of large deviations through a simple coin tossing example, introducing the idea of decay rate functions and limit driven probability bounds. It will then build successively tighter bounds on large deviation decay rate functions for iid random variables through proof-backed theorems, culminating in Cramér's Theorem for the empirical average. A thorough framework for introductory large deviation theory will be constructed through these theorems. The next layers of complexity naturally following these introductory conclusions will be alluded to, and we will use our LDT tools to examine familiar distributions of random variables.

Definitions and Preliminary Theory

To properly introduce large deviation theory, we must first standardize the notation for basic elements of probability theory. For now, we will restrict our scope to iid random variables defined on a probability space $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$. For a sequence $(X_i)_{i\in\mathbb{N}}$ of iid $\mathbb{R}$-valued random variables, let $S_n = \sum_{i=1}^{n} X_i$ be the partial sums of X_i. Every LDT theorem we prove will involve the limits of these partial sums. Let the expectation value and variance of X_1 be defined as

$$\mathbb{E}[X_1] = \mu \in \mathbb{R} \quad \text{and} \quad \operatorname{Var}(X_1) = \sigma^2 \in (0, \infty) \tag{1}$$

respectively. By the iid assumption, these values of μ and σ^2 apply to all X_i. Now, from probability theory, we know the following:

Strong Law of Large Numbers (SLLN)

$$\frac{1}{n}S_n \xrightarrow[n\to\infty]{} \mu \quad \mathbb{P}\text{-a.s.} \tag{2}$$

Central Limit Theorem (CLT)

$$\frac{1}{\sigma\sqrt{n}}(S_n - \mu n) \xrightarrow[n\to\infty]{} Z \quad \text{in law w.r.t } \mathbb{P} \tag{3}$$

The SLLN posits that the arithmetic mean of a sequence of iid RVs approaches the expectation value of X_i almost surely as $n \to \infty$ (with probability 1). The CLT conveys the property that the sum of iid RVs approaches a normal distribution as $n \to \infty$. Engrained within the CLT is the idea of a "normal deviation", which we will explore shortly.

A popular inequality for imposing bounds on probabilities is Markov's inequality,

$$\mathbb{P}(X_1 \geq a) \leq \frac{\mathbb{E}[X_1]}{a}, \qquad \text{with } X_1 \geq 0, \tag{4}$$

which we will use in a LDT proof. Another tool that we will use is Chebyshev's Inequality,

$$\mathbb{P}\left(|X_1 - \mu| \geq c\right) \leq \frac{\mathrm{Var}(X_1)}{c^2}, \tag{5}$$

which will prove useful in proving large deviation bounds.

Moment-Generating Functions

We will later see that moment-generating functions of random variables are of utmost importance in calculating rate functions for large deviation probabilities. Thus, it is necessary to have a working knowledge of mathematical moments and moment-generating functions of random variable distributions.

If X is a continuous random variable defined on the probability space Ω, with a probability density function f_X, then we define the nth moment of X by the formula

$$\mu_n = \mathbb{E}[X^n] = \int_{-\infty}^{+\infty} x^n f_X(x) dx, \tag{6}$$

provided the integral

$$\mu_n = \mathbb{E}[X^n] = \int_{-\infty}^{+\infty} |x|^n f_X(x) dx$$

is finite. We can see through observation that $\mu_0 = 1$, $\mu_1 = \mu$, and $\mu_2 - \mu_1^2 = \sigma^2$. It turns out that if the series converges on an open interval containing the origin, then the moments uniquely determine the distribution. Now, let us define the moment-generating function $\varphi(t)$ for X by the formula

$$\begin{aligned} \varphi(t) = \mathbb{E}[e^{tX}] &= \int_{-\infty}^{+\infty} e^{tx} f_X(x) dx \\ &= \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mathbb{E}[X^k] t^k}{k!} \end{aligned} \tag{7}$$

provided this series converges. Now, we observe that

$$\left. \frac{d^n}{dt^n} \varphi(t) \right|_{t=0} = \varphi^{(n)}(0) = \left. \sum_{k=n}^{\infty} \frac{k! \mu_k t^{k-n}}{(k-n)! k!} \right|_{t=0} = \mu_n \ ,$$

so we have the relation,

$$\mu_n = \varphi^{(n)}(0). \tag{8}$$

This equation states that to retrieve the nth moment of a random variable from its moment-generating function, we compute the nth derivative of $\varphi(t)$ evaluated at $t = 0$. Thus, the moment-generating function is a compact function containing a probability distribution's information – in fact, for bounded, real valued random variables, $\varphi(t)$ determines $f_X(x)$ uniquely. This concept of moment-generating functions applies to discrete random variables with

$$\mu_n = \mathbb{E}\left[X^n\right] = \sum_{j=1}^{\infty} (x_j)^k p(x_j),$$

where $p(x_j) = \mathbb{P}(X = x_j)$, and

$$\varphi(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{j=1}^{\infty} e^{t x_j} p(x_j) \ .$$

We will find that under certain conditions of a random variable's moment-generating function, we will be able to directly determine the probabilistic rate function of large deviations. The discussion in this section follows from Grinstead's *Introduction to Probability*.[2]

To familiarize ourselves with moment-generating functions, lets compute $\varphi(t)$ for a handful of familiar iid random variables.

Binomial Distribution:

Let X be a discrete random variable such that $X \sim \text{Binom}(n, p)$ with probability density function $p_X(j) = \binom{n}{j} p^j q^{n-j}$ for $0 \leq j \leq n$. Now, we compute the moment-generating function using our definition for discrete random variables as follows:

$$\begin{aligned} \varphi(t) &= \sum_{j=0}^{n} e^{tj} \binom{n}{j} p^j q^{n-j} \\ &= \sum_{j=0}^{n} \binom{n}{j} (pe^t)^j q^{n-j} \\ &= (pe^t + q)^n \ . \end{aligned} \tag{9}$$

We can easily check that

$$\begin{aligned} \mu_1 &= \varphi'(0) = n(pe^t + q)^{n-1} pe^t|_{t=0} = np \\ \mu_2 &= \varphi''(0) = np((n-1)p + 1) \ , \end{aligned}$$

so that $\mathbb{E}[X] = \mu_1 = np$ and $\text{Var}(X) = \mu_2 - \mu_1^2 = np(1-p)$, matching our expressions for the expected value and variance of random variables with binomial distributions.

Normal Distribution:

Let X be a continuous random variable on $(-\infty, +\infty)$ such that $X \sim N(\mu, \sigma^2)$, so we have

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \ .$$

Calculating the moments directly leads to tricky integrals, but we know that $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Now, to find the moment-generating function,

$$\varphi(t) = \mathbb{E}[e^{xt}] = \int_{\infty}^{+\infty} e^{xt} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \ .$$

Now, let $z = \frac{x-\mu}{\sigma}$, so $x = z\sigma + \mu$. Using this change of variables we have

$$\begin{aligned}
\varphi(t) &= e^{\mu t} \int_{\infty}^{+\infty} e^{z\sigma t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2} \left|\frac{dx}{dz}\right| dz \\
&= e^{\mu t} \int_{\infty}^{+\infty} e^{z\sigma t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \qquad \text{as } \frac{dx}{dz} = \sigma \\
&= e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} \\
&= e^{\mu t + \frac{\sigma^2 t^2}{2}} \ .
\end{aligned} \tag{10}$$

Now we can check that

$$\begin{aligned}
\mu_1 &= \varphi'(0) = (\mu + \sigma^2 t) e^{\mu t + \frac{\sigma^2 t^2}{2}}|_{t=0} = \mu \\
\mu_2 &= \varphi''(0) = (\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 + \sigma^2) e^{\mu t + \frac{\sigma^2 t^2}{2}}|_{t=0} = \mu^2 + \sigma^2 \ ,
\end{aligned}$$

so that $\mathbb{E}[X] = \mu_1 = \mu$ and $\mathrm{Var}(X) = \mu_2 - \mu_1^2 = (\mu^2 + \sigma^2) - \mu^2 = \sigma^2$, matching our expressions for the expected value and variance of random variables with normal distributions.

Exponential Distribution:

Let X be a continuous random variable on $[0, \infty)$ such that $X \sim \mathrm{Expo}(\lambda)$, so that $f_X(x) = \lambda e^{-\lambda x}$. Here,

$$\mu_n = \int_0^{\infty} x^n \lambda e^{-\lambda x} dx = \lambda(-1)^n \frac{\partial^n}{\partial \lambda^n} \int_0^{\infty} e^{-\lambda x} dx = \lambda(-1)^n \frac{\partial^n}{\partial \lambda^n} \frac{1}{\lambda} = \frac{n!}{\lambda^n} \ ,$$

so we have

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!} = \sum_{k=0}^{\infty} \left(\frac{t}{\lambda}\right)^k = \frac{\lambda}{\lambda - t} \ , \tag{11}$$

which converges for $t < \lambda$. Now we can easily check that

$$\begin{aligned}
\mu_1 &= \varphi'(0) = \frac{\lambda}{(\lambda - t)^n}|_{t=0} = \frac{1}{\lambda} \\
\mu_2 &= \varphi''(0) = \frac{2\lambda}{(\lambda - t)^3}|_{t=0} = \frac{2}{\lambda^2} \ ,
\end{aligned}$$

so that $\mathbb{E}[X] = \mu_1 = 1/\lambda$ and $\mathrm{Var}(X) = \mu_2 - \mu_1^2 = 1/\lambda^2$, matching our expressions for the expected value and variance of random variables with exponential distributions.

II. LARGE DEVIATIONS AND RATE FUNCTIONS

Now let's begin to consider large deviations and their inherent properties. We can see that the *CLT* accounts for deviations of S_n from μn of order $\sqrt{n}$, describing central tendencies in a distribution. These are typical deviations, where the *CLT* and *SLLN* apply. Large deviation theory on the other hand explores the probability that S_n differs from μn by an amount of order n. While the probability of an event such as $\{S_n \geq an\}$ with $a > \mu$ tends to zero as $n \to \infty$, we must consider the rate at which it converges to zero. We find that under certain conditions of the tail of the distribution of X_1, the decay is exponential in n. This result is not intuitive, and its significance is lost without an investigation into the underlying theory. To introduce the framework of LDT, let us first observe large deviations through a classic example.

Coin Tossing Example

Perhaps the best introduction to LDT is through one of the most basic examples in probability theory – the simple coin toss. In this example we observe a sequence of iid random variables (X_i) with $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = \frac{1}{2}$. For $a,\ b \in (0,1)$ and $a < b$, we know that the probability of the average number of heads in a certain range is

$$\mathbb{P}(a < S_n/n < b) = \sum_{\{k\,:\,a<\frac{k}{n}<b\}} \binom{n}{k} \frac{1}{2^n} \tag{12}$$

where k is the total number of heads, and n is the total number of tosses. This is the sum of the discrete probabilities for each possible number of heads in the given range.

To begin understanding the probabilistic behavior of this coin tossing example, let's observe the distribution of S_n/n for increasing values of n. It is clear from Fig 1. that the SLLN is at work here – as the number of coin tosses n increases, $\frac{1}{n}S_n$ approaches μ, and the tails of the distribution become smaller. To quantify the shrinking of these tails, let's first observe the probability that the average number of heads per coin toss is greater than 0.6 for increasing n. In observing $\mathbb{P}\left(S_n/n > 0.6\right)$ and $\log \mathbb{P}\left(S_n/n > 0.6\right)$ in Fig 2., we can see that there is clear exponential behavior.

Now, by graphing the logarithm of these probabilities for lower bounds greater than 0.6 in Fig. 3, we see the same behavior. The tail of the distribution of the average number of

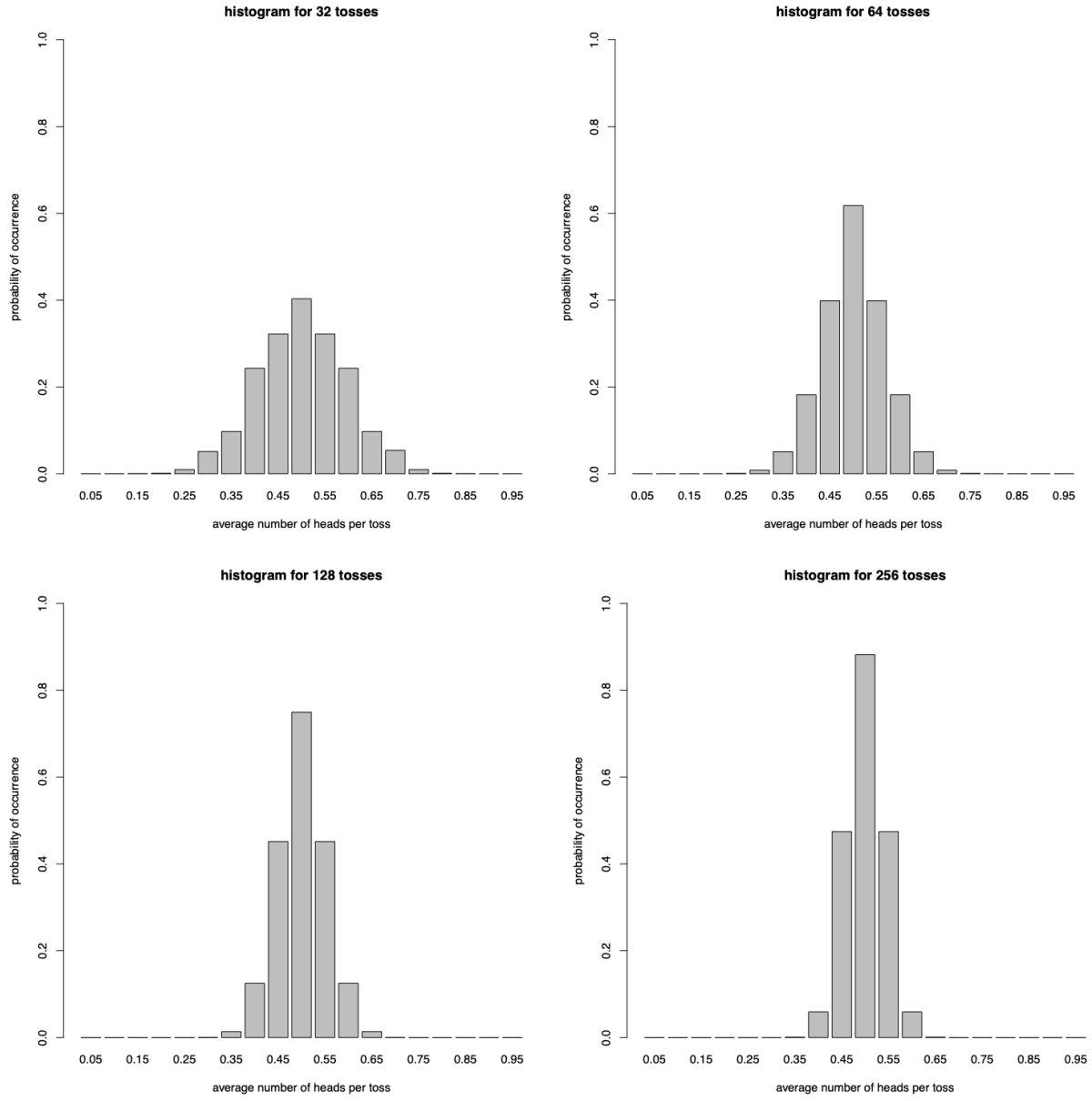

FIG. 1. Probability distribution of average number of heads per toss in coin tossing example.

heads in n tosses decays exponentially as n increases. In all of these graphs, we see a log-linear relationship with increasing n, and increasingly negative slopes with higher values of a. While there is variation in the slope for low n, we see an asymptotic slope as n increases. Let us refer to the asymptotic slope of these lines as $-I(a)$. Thus, we expect the probability of the average number of heads to be a function with the form

$$\mathbb{P}(S_n/n \geq a) \approx e^{-nI(a)},$$

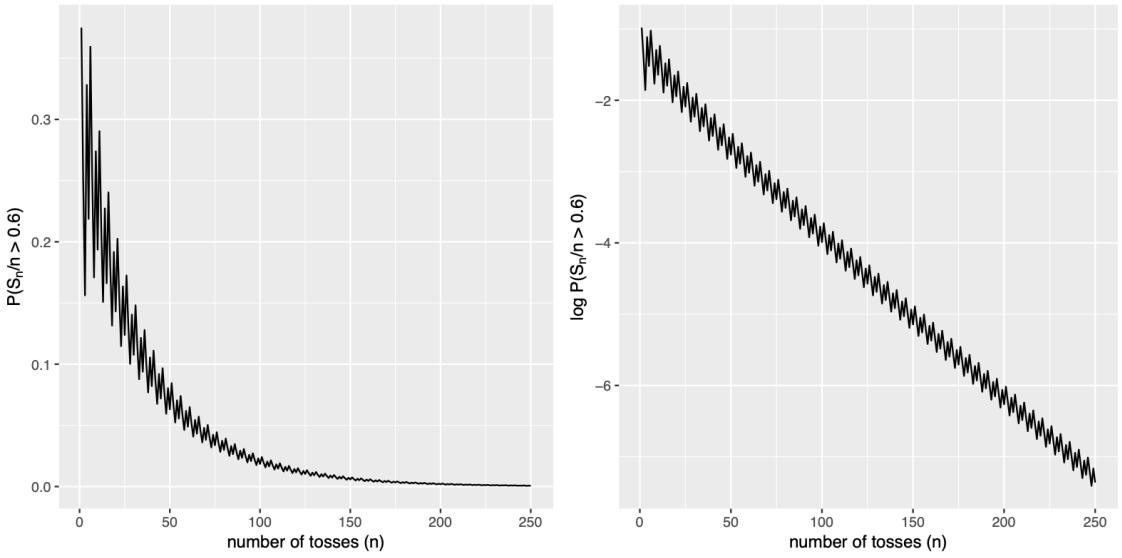

FIG. 2. $\mathbb{P}\left(S_n/n > 0.6\right)$ and $\log \mathbb{P}\left(S_n/n > 0.6\right)$ against n.

or, restated by taking the logarithm:

$$\log \mathbb{P}(S_n/n \geq a) \approx -nI(a),$$

thus the behavior of these probability distribution tails lies in the "logarithmic slope function", $I(a)$. But how do we find this function?

Now, we claim that in our coin flipping example, for all $a > \frac{1}{2}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n/n \geq a) = -I(a), \tag{13}$$

where

$$I(z) = \begin{cases} \log 2 + z \log z + (1-z)\log(1-z) & \text{if } z \in [0,1], \\ \infty & \text{otherwise.} \end{cases} \tag{14}$$

To prove this, let's remind ourselves that $\mathbb{P}(a < S_n/n < b) = \frac{1}{2^n} \sum_{\{k\,:\, a<\frac{k}{n}<b\}} \binom{n}{r}$. To account only for the lower tail of the distribution, we can restate this as $\mathbb{P}(S_n/n < a) = \frac{1}{2^n} \sum_{k=0}^{\lceil a\rceil - 1} \binom{n}{k}$, where $\lceil a \rceil$ is the smallest integer greater than or equal to a. Now, if $a < \frac{1}{2}$, then we know each individual term in the sum is bounded by $\binom{n}{\lceil na\rceil}$. Now, we have

$$\mathbb{P}(S_n/n < a) \leq \lceil na \rceil \binom{n}{\lceil na \rceil} \frac{1}{2^n},$$

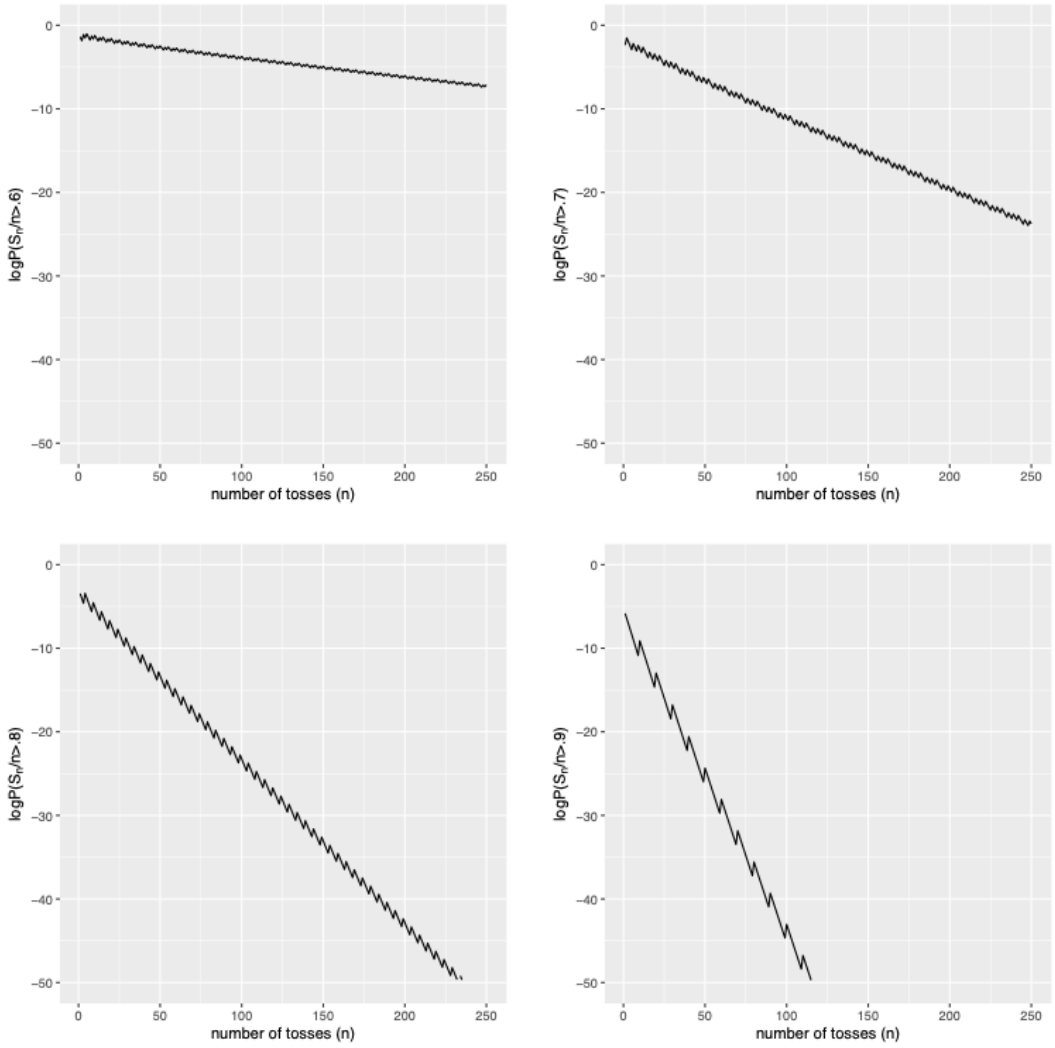

FIG. 3. $\mathbb{P}\left(S_n/n > 0.6\right)$ and $\log \mathbb{P}\left(S_n/n > 0.6\right)$ against n.

which we will now refer to as U_n, an upper bound for this probability. Let us now consider $\log U_n$. We can expand $\log \binom{n}{\lceil na \rceil}$ to

$$\begin{aligned}\log \binom{n}{\lceil na \rceil} = &-\frac{\lceil na \rceil}{n}\left(\frac{\log \lceil na \rceil!}{\lceil na \rceil} - \log \lceil na \rceil\right) \\ &- \frac{\lfloor n(1-a) \rfloor}{n}\left(\frac{\log \lfloor n(1-a) \rfloor!}{\lfloor n(1-a)} - \log \lfloor n(1-a) \rfloor\right) + \left(\frac{\log n!}{n} - \log n\right) \\ &- \frac{\lceil na \rceil}{n} \log \frac{\lceil na \rceil}{n} - \frac{\lfloor n(1-a) \rfloor}{n} \log \frac{\lfloor n(1-a) \rfloor}{n},\end{aligned}$$

and by using Stirling's Formula,

$$\log n! = n \log n - n + \mathcal{O}(\log n) \tag{15}$$

we can see that the expression

$$\lim_{n\to\infty}\left(\frac{1}{n}\log n! - \log n\right) = -1\ , \tag{16}$$

is true. Now with this expression, we arrive at

$$\lim_{n\to\infty}\frac{1}{n}\log U_n = -\log 2 - a\log a - (1-a)\log(1-a),$$

thus we have proven (14) for the upper bound of U_n. Now, we can see that for $a > 0$,

$$\mathbb{P}(S_n/n < a) \geq \binom{n}{\lceil na\rceil - 1}\frac{1}{2^n},$$

which we will call L_n similarly to our upper bound U_n as before. It can be shown by a similar procedure that

$$\lim_{n\to\infty}\frac{1}{n}\log L_n = -\log 2 - a\log a - (1-a)\log(1-a),$$

which we will not directly show in this paper. Thus, by showing that the upper and lower bounds converge to the same function, $-I(a)$, we have proven (14) to be true. We can see how $I(a)$ changes for varying probabilities in Fig 4. The discussion in this section follows from Lewis and Russel's *Introduction to Large Deviations for Teletraffic Engineers*.[5]

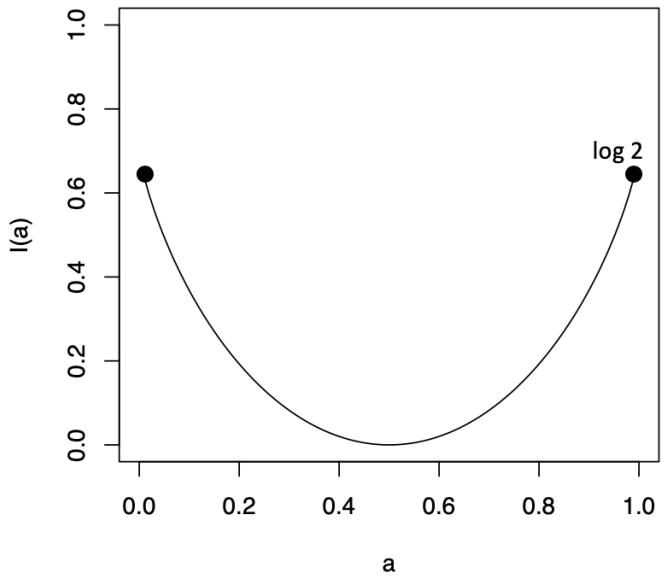

FIG. 4. $I(a)$ for the coin flipping example.

Large Deviation Principle

We found in our coin tossing example that the probabilistic decay rate for large deviations had exponential behavior with increasing n. It turns out that this relationship is not uncommon – in fact, it is the basic principle of large deviations for simple iid random variables. The behavior of coin flipping example evokes what we refer to as the Large Deviation Principle.

Large Deviation Principle (LDP) For a sequence of iid random variables (X_i), if the moment-generating function for X_1 has the property $\varphi(t) < \infty \; \forall \; t$ in some open neighbourhood of 0, then for large n and some $a > \mu$, there exists a rate function I such that

$$\mathbb{P}(S_n/n \geq a) \approx e^{-nI(a)} \tag{17}$$

Or, restated:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \geq an\big) = -I(a) < 0, \quad a > 0. \tag{18}$$

Here, $-I(a)$ is the rate at which the probability of large deviations in sequence of partial sums S_n goes to zero as n goes to infinity. From here on out, we will mainly be looking at large deviations by observing these rate functions. We will look at examples beyond the simple coin tosses, and build theorems to tighten bounds and bolster our understanding.

Chernoff Bound

To begin defining a procedure for determining rate functions, lets find a suitable upper bound. Fixing a positive parameter $t > 0$, we have

$$\begin{aligned}
\mathbb{P}(S_n > na) &= \mathbb{P}\left(e^{tS_n} > e^{tna}\right) && e^x \text{ is monotone increasing} \\
&\leq \frac{\mathbb{E}\left[e^{tS_n}\right]}{e^{tna}} && \text{Markov inequality} \\
&= \frac{\mathbb{E}\left[\prod_i e^{tX_i}\right]}{(e^{ta})^n} && X_i\text{'s independent} \\
&= \left(\frac{\mathbb{E}[e^{tX_1}]}{e^{ta}}\right)^n && X_i\text{'s are identically distributed} \\
&= \left(\frac{\varphi(t)}{e^{ta}}\right)^n && \text{moment-generating function of } X_1.
\end{aligned}$$

By optimizing t to minimize this expression, this inequality is known as the Chernoff Bound,

$$\mathbb{P}(S_n > na) \leq \inf_{t>0} \left(\frac{\varphi(t)}{e^{ta}} \right)^n, \tag{19}$$

casting an exponentially decreasing bound on the tail distributions of sums of independent random variables. This derivation of the Chernoff bound follows from the MIT *Large Deviations for iid Random Variables* lecture notes.[6] This provides a tighter bound than the Markov or Chebyshev inequalities alone. Now, by massaging this bound into the form of the LDP,

$$\begin{aligned}
\mathbb{P}(S_n > na) &\leq \inf_{t>0} \left(\frac{\varphi(t)}{e^{ta}} \right)^n \\
\log \mathbb{P}(S_n > na) &\leq \inf_{t>0} n \log \left(\frac{\varphi(t)}{e^{ta}} \right) \qquad \text{as } \log \text{ is monotone increasing} \\
\frac{1}{n} \log \mathbb{P}(S_n > na) &\leq \inf_{t>0} [\log \varphi(t) - ta] \\
\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > na) &\leq \inf_{t>0} [\log \varphi(t) - ta] \qquad \text{as RHS does not rely on } n.
\end{aligned}$$

We can freely pass a logarithm through an infimum because as log is a monotone increasing function, all inequalities in infimums and supremums still holds. Now since $\sup[-f(x)] = -\inf[f(x)]$, from the expression for the rate function $I(a)$ from the LDP, we have

$$I(a) \leq \sup_{t>0} [at - \log \varphi(t)] \quad . \tag{20}$$

This is our first definite result in determining rate function for large deviations of iid random variables. In essence, it states that the upper bound of the rate function is determined by the magnitude of the deviation, a, and the moment-generating function parameter, t. In the next section we will discover that this is not an extraneous bound – in fact, for an optimized parameter t, the Chernoff bound *defines* the rate function.

Transforms and Cramér's Theorem

Our expression in (20) bounding the rate function is what we call a Fenchel-Legendre transform. The Fenchel-Legendre transform $\mathcal{F}[f] : \mathcal{I}^* \mapsto \mathbb{R}$ of a function $f(x)$ has the form

$$\mathcal{F}[f] \stackrel{\text{def}}{=} \sup_{x\in\mathcal{I}} [x^* x - f(x)], \quad x^* \in \mathcal{I}^*,$$

where the domain $\mathcal{I}^*$ is

$$\mathcal{I}^* = \{x^* \in \mathbb{R} : \sup_{x \in \mathcal{I}}[x^* x - f(x)] < \infty\}.$$

This transform is useful in describing rate functions as it is always well-defined for convex functions, and we will shortly see that $\varphi(t)$ is a strictly convex function. Another transform that we will shortly utilize is the Laplace transform, which has the form

$$\mathcal{L}[f](\lambda) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} e^{-\lambda x} f(x) dx, \quad \lambda \in \mathcal{D},$$

where the domain $\mathcal{D}$ is

$$\mathcal{D} = \{\lambda \in \mathbb{R} : \int_{-\infty}^{+\infty} e^{-\lambda x} f(x) dx < \infty\}.$$

The Laplace transform transforms differential equations into algebraic equations and convolution into multiplication. Finally, the Cramér transform $f \mapsto \mathcal{C}[f] \stackrel{\text{def}}{=} \mathcal{F}[\log \mathcal{L}[f]]$ is the Legendre-Fenchel transform applied to the logarithm of the Laplace transform. Cramér transforms are useful for mapping Gaussians into corresponding quadratics. We will use Cramér transforms shortly. The discussion on transformations follows from Lasserre's *Linear and Integer Programming vs Linear Integration and Counting*.[4]

Now, after defining these transforms, let's use the Chernoff bound to uniquely determine rate functions through Cramér's Theorem.

Cramér's Theorem Let (X_i) be a sequence of iid $\mathbb{R}$–valued random variables satisfying

$$\varphi(t) = \mathbb{E}\left[e^{tX_1}\right] < \infty \quad \forall t \in \mathbb{R}. \tag{21}$$

Let $S_n = \sum_{i=1}^n X_i$. Then for all $a > \mathbb{E}[X_1]$,

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq an) = -I(a), \tag{22}$$

where

$$I(z) = \sup_{t \in \mathbb{R}}[zt - \log \varphi(t)]. \tag{23}$$

In proving this theorem, we can assume without loss of generality that $a = 0$ and $\mathbb{E}[X_1] < 0$ so that we are observing $I(0)$, simplifying our expression of the Legendre transform. Thus,

we are trying to prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) = -I(0) = \sup_{t\in\mathbb{R}}[\log \varphi(t)]$$

From here on, as $\sup[-f(x)] = \inf[f(x)]$, let $\rho = \inf_{t\in\mathbb{R}}[\varphi(t)] = \sup_{t\in\mathbb{R}}[-\varphi(t)]$. Restating (II), we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) = \log \rho \quad . \tag{24}$$

Now in observing the moment generating function $\varphi(t)$ and the cumulative distribution function $F(x) = \mathbb{P}(X_1 \leq x)$ for X_1, following (21), we know that

$$\varphi'(t) = \int_{\mathbb{R}} x e^{tx} dF(x)$$
$$\varphi''(t) = \int_{\mathbb{R}} x^2 e^{tx} dF(x)$$

Now, as x^2 is an even function, we know that $\varphi''(t) \geq 0$, thus $\varphi(t)$ is a strictly convex function. Additionally, we can see that $\varphi'(0) = \mathbb{E}[X_1] < 0$. Now, to prove (24), lets divide the probability distribution of X into three cases:

(i) $\mathbb{P}(X_1 < 0) = 1$.

In this case, in observing $\varphi'(t)$, we can see that $\varphi(t)$ is strictly decreasing. Thus $\lim_{t\to\infty} \varphi(t) = \rho = 0$. Additionally, $\mathbb{P}(S_n \geq 0) = 0$, so by substituting these values into (24), we arrive at $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) = \log \rho$, and the claim is supported.

(ii) $\mathbb{P}(X_1 \leq 0) = 1$ and $\mathbb{P}(X_1 = 0) > 0$.

Again, here we see that $\varphi(t)$ is strictly decreasing, and that $\lim_{t\to\infty} \varphi(t) = \rho = \mathbb{P}(X_1 = 0) > 0$. Now, we observe that $\mathbb{P}(S_n \geq 0) = \mathbb{P}(X_1 = ... = X_n = 0) = \rho^n$. With this in mind, we see that

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) &= \lim_{n\to\infty} \frac{1}{n} \log \rho^n \\ &= \lim_{n\to\infty} \frac{1}{n} n \log \rho \\ &= \log \rho \quad , \end{aligned}$$

thus the claim is supported.

(iii) $\mathbb{P}(X_1 < 0) > 0$ and $\mathbb{P}(X_1 > 0) > 0$.

Now here, we observe that $\lim_{t\to\infty} \varphi(t) = \infty$, and since $\varphi(t)$ is strictly convex, there must exist a unique $\tau \in \mathbb{R}$ satisfying $\tau > 0$ s.t. $\varphi(\tau) = \rho$ and $\varphi'(\tau) = 0$. In other words, this optimized value τ minimizes φ.

Now to prove (24) for this case, we will show that the supremum and the infimum of the expression are equal. First, we observe that

$$\begin{aligned}\mathbb{P}(S_n \geq 0) &\leq \left(\frac{\varphi(\tau)}{e^{\tau 0}}\right)^n \qquad \text{Chernoff Bound} \\ &= (\varphi(\tau))^n \\ &= \rho^n \,.\end{aligned}$$

From this relationship, we see that

$$\lim_{n\to\infty} \sup \frac{1}{n} \log \mathbb{P}(S_n \geq 0) \leq \log \rho \,, \tag{25}$$

defining the upper bound as $\log \rho$.

Now, we must find a way to determine the lower bound. We will employ a technique using Cramér transforms. Let $(\widehat{X}_i)$ be iid sequence of random variables with distribution function

$$\widehat{F}(x) = \frac{1}{p} \int_{(-\infty,x]} e^{\tau y} dF(y) \tag{26}$$

where $\widehat{F}(x)$ is the Cramér transform of $F(x)$, and $\rho = \varphi(\tau) = \int_{\mathbb{R}} e^{\tau y} dF(y)$ as defined before. We will now use the Cramér transform to prove the lower bound through a series of Lemmas.

Lemma I $\mathbb{E}[\widehat{X}_1] = \widehat{\mu} = 0$ and $Var(\widehat{X}_1) = \widehat{\sigma}^2 \in (0, \infty)$

Proof. Let $\widehat{\varphi}(t) = \mathbb{E}[e^{t\widehat{X}_1}]$. Additionally, from (26), we have $d\widehat{F}(x) = \frac{1}{\rho} e^{\tau x} dF(x)$ Now,

$$\begin{aligned}\widehat{\varphi}(t) &= \int_{\mathbb{R}} e^{tx} d\widehat{F}(x) = \frac{1}{\rho} \int_{\mathbb{R}} e^{tx} e^{\tau x} dF(x) \\ &= \frac{1}{\rho} \int_{\mathbb{R}} e^{x(t+\tau)} dF(x) = \frac{1}{\rho} \varphi(t+\tau) = \frac{\varphi(t+\tau)}{\varphi(\tau)},\end{aligned}$$

and since $\varphi(t) < \infty \; \forall t \in \mathbb{R}$ from (21) and $\varphi(\tau) > 0$, $\frac{\varphi(t+\tau)}{\varphi(\tau)} < \infty \; \forall t \in \mathbb{R}$. This implies that $\widehat{\varphi}$ is infinitely differentiable on $\mathbb{R}$, so

$$\mathbb{E}[\widehat{X}_1] = \widehat{\varphi}'(0) = \frac{1}{\rho} \frac{\partial}{\partial t} \left(\frac{\varphi(t+\tau)}{\varphi(\tau)}\right) = \frac{\varphi'(\tau)}{\varphi(\tau)},$$

and since $\varphi'(\tau) = 0$ as defined before, $\mathbb{E}[\widehat{X}_1] = 0$.

Similarly,

$$\mathrm{Var}(\widehat{X}_1) = \widehat{\varphi}''(0) = \frac{1}{\rho}\varphi''(\tau) \in (0, \infty)$$

as we know that φ is strictly convex, so $\varphi'' > 0$. □

Lemma II Let $\widehat{S}_n = \sum_{i=1}^{n} \widehat{X}_1$. Then $\mathbb{P}(S_n \geq 0) = \rho^n \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right]$

Proof. Now, using the fact that $dF(x) = \rho e^{-\tau x} d\widehat{F}(x)$, we have

$$\begin{aligned}
\mathbb{P}(S_n \geq 0) = \mathbb{P}((X_1 + ... + X_n) \geq 0) &= \int\limits_{\{x_1+...+x_n \geq 0\}} dF(x_1)...dF(x_n) \\
&= \int\limits_{\{x_1+...+x_n \geq 0\}} \left[\rho e^{-\tau x_1} d\widehat{F}(x_1)\right] ... \left[\rho e^{-\tau x_n} d\widehat{F}(x_n)\right] \\
&= \rho^n \int\limits_{\{x_1+...+x_n \geq 0\}} e^{-\tau(x_1+...+x_n)} d\widehat{F}(x_1)...d\widehat{F}(x_n) \\
&= \rho^n \int\limits_{\{x_1+...+x_n \geq 0\}} e^{-\tau S_n} d\widehat{F}(x_1)...d\widehat{F}(x_n) \\
&= \rho^n \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] .
\end{aligned}$$

□

Lemma III $\liminf\limits_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] \geq 0$

Proof. Using Lemma I, we observe that the CLT can be applied to $\widehat{S}_n$, thus $\frac{\widehat{S}_n - n\widehat{\mu}}{\widehat{\sigma}\sqrt{n}} = \frac{\widehat{S}_n}{\widehat{\sigma}\sqrt{n}} \to Z$. Additionally, let $C > 0$ be any number such that $\frac{1}{\sqrt{2\pi}} \int\limits_0^C e^{-x^2/2} dx > \frac{1}{4}$. Now to simplify the ensuing proof, let us split the indicator $\mathbb{1}_{\{\widehat{S}_n \geq 0\}}$ into indicators for two disjoint events,

$$\begin{aligned}
\mathbb{1}_{\{\widehat{S}_n \geq 0\}} &= \mathbb{1}_{\{0 \leq \widehat{S}_n < C\widehat{\sigma}\sqrt{n}\}} + \mathbb{1}_{\{\widehat{S}_n \geq C\widehat{\sigma}\sqrt{n}\}} \\
&= \mathbb{1}_A + \mathbb{1}_B \qquad \text{to ease notation.}
\end{aligned}$$

Now,

$$\mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] = \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_A\right] + \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_B\right] \qquad \text{additivity of } \mathbb{E}$$

and since $\mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_B\right] \geq 0$ and the minimum of $e^{-\tau \widehat{S}_n}$ on $(0 \leq \widehat{S}_n < C\widehat{\sigma}\sqrt{n})$ is $e^{\tau C\widehat{\sigma}\sqrt{n}}$, we have

$$\mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] \geq \left(e^{\tau C\widehat{\sigma}\sqrt{n}}\right) \mathbb{E}[\mathbb{1}_A] = \left(e^{\tau C\widehat{\sigma}\sqrt{n}}\right) \mathbb{P}(A).$$

Now, by our choice of C, we know $\mathbb{P}(A) = \mathbb{P}(0 \leq \frac{\widehat{S}_n}{\widehat{\sigma}\sqrt{n}} < C) > \frac{1}{4}$ for n large as $\widehat{S}_n/\sqrt{n} \to Z$. Thus, we have

$$\begin{aligned}
\log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] &\geq \log\left[\left(e^{\tau C \widehat{\sigma}\sqrt{n}}\right)\frac{1}{4}\right] \\
\frac{1}{n}\log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] &\geq \frac{\tau C \widehat{\sigma}\sqrt{n} + \log(\frac{1}{4})}{n} \\
\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] &\geq \liminf_{n\to\infty} \frac{\tau C \widehat{\sigma}\sqrt{n} + \log(\frac{1}{4})}{n} \\
\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(S_n \geq na) &\geq 0 \ .
\end{aligned}$$

□

Now, using Lemmas II and III, we observe

$$\begin{aligned}
\mathbb{P}(S_n \geq 0) &= \rho^n \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] \qquad \text{(Lemma II)} \\
\log \mathbb{P}(S_n \geq 0) &= n \log \rho + \log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] \\
\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(S_n \geq 0) &= \log \rho + \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{1}_{\{\widehat{S}_n \geq 0\}}\right] \\
\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(S_n \geq 0) &\geq \log \rho \qquad \text{as 2nd term} \geq 0 \text{ (Lemma III)},
\end{aligned}$$

thus proving the lower bound to be $\log \rho$. We have now shown that the tight upper and lower bounds are both $\log \rho$, so the claim for case (iii) is true, effectively proving Cramér's Theorem for the empirical average. It also immediately follows that Cramér's Theorem holds for $\mathbb{P}(S_n \leq an)$ where $a < \mathbb{E}[X_1]$ with the same formula for I. Thus, it applies to large deviations on both sides of a partial sum's expected value. The discussion in this section follows den Hollander's *Large Deviations*.[3]

We now have a widely applicable approach for directly determining the rate functions for large deviations of iid random variables. We found that this rate function is simply the Legendre-Fenchel transform of the logarithm of a random variable's moment generating function. The only requirement for this approach is that the moment generating function $\varphi(t)$ be bounded for all $t \in \mathbb{R}$. While there exist generalizations loosening this requirement, they require tools in probability theory and set topology outside the scope of this paper. Yet one important generalization weakens the condition on $\varphi(t)$ to

$$0 \in \text{int}(\mathcal{D}_\varphi) \qquad \text{with} \qquad \mathcal{D}_\varphi = \{t \in \mathbb{R} : \varphi(t) < \infty\} \tag{27}$$

where $\text{int}(\mathcal{D}_\varphi)$ is the interior of the set $\mathcal{D}_\varphi$. This states that for Cramér's Theorem to apply, one must only make sure that $\varphi(t) < \infty \; \forall \; t$ in some open neighborhood of 0. This is the same initial condition for the earlier stated *LDP*. The proof of this condition can be found in Billingsley's *Probability and Measure*.[1] While we will not prove this generalization in this paper, it is reasonably intuitive and will be useful in computing rate functions when the moment generating functions for iid random variables are not everywhere bounded on $\mathbb{R}$.

III. RATE FUNCTION CALCULATIONS

Let us now examine the large deviation behavior of a handful of iid random variable distributions by calculating their consequent rate functions, $I(t)$. We will begin by revisiting our coin flipping example.

Binomial Distribution:

Let (X_n) be a sequence of iid discrete random variables such that $X_1 \sim \text{Binom}(1, p)$. This variable represents our coin flipping example, but allows asymmetric probabilities, modeling a "biased" coin. We remind ourselves that $\varphi(t) = (pe^t + q)^n = (pe^t + 1 - p)$. Evoking the generalization stated in (27), we can see that $\varphi(0 \pm \epsilon) < \infty$, so $0 \in \mathcal{D}_\varphi$, satisfying the initial condition for Cramér's Theorem. Thus, the rate function is given by

$$I(z) = \sup_{t \in \mathbb{R}}[zt - \log \varphi(t)] = \sup_{t \in \mathbb{R}}[zt - \log(pe^t + 1 - p)] \; .$$

Now to find the supremum of this expression, τ, let's take the first derivative with respect to t,

$$\frac{\partial}{\partial t}(zt - \log(pe^t + 1 - p)) = z - \frac{pe^t}{pe^t + 1 - p} \; ,$$

and by setting this expression equal to zero, we find that the value of t that maximizes the expression is

$$\tau = \log \frac{z}{p} + (1 - z) \log \frac{1 - z}{1 - p} \; .$$

Thus, the rate function for a series of iid RVs with Binomial distributions is

$$I(z) = z \log \frac{z}{p} - (1 - z) \log \frac{1 - z}{1 - p} \; .$$

To model a "fair" coin toss, we use a value of $p = 0.5$, giving us the familiar rate function, $I(z) = \log 2 + z \log z + (1 - z) \log(1 - z)$ for $z \in [0, 1]$. Our result using Cramér's Theorem

matches that using combinatorics! We can see that for values of z further away from p, the rate function increases. This behavior agrees with our earlier log plots, where the slope was more negative for increasing z. To emphasize the power of our new generalized rate function, let's observe the exponential tail decay rate for a value of $p = 0.25$ in Fig 5.

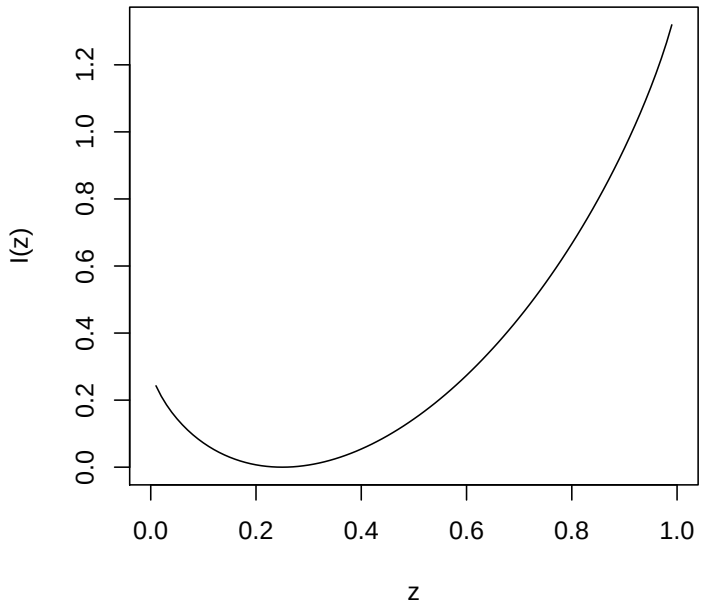

FIG. 5. The rate function for coin tossing with a biased coin (p = 0.25).

We can see that $I(0.25) = 0$ as $\mathbb{E}[S_n/n] = .25 = \mathbb{E}[X_1]$, and the exponential tail decay rate increases for larger deviations from the expected value. Additionally, $I(z) = \infty \quad \forall z \notin [0, 1]$.

Normal Distribution:

Now, let (X_n) be a sequence of iid random variables such that $X_1 \sim N(\mu, \sigma^2)$. We showed earlier that $\varphi(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$. Evoking the generalization stated in (27), we can see that $\varphi(0 \pm \epsilon) < \infty$, so $0 \in \mathcal{D}_\varphi$, satisfying the initial condition for Cramér's Theorem. Thus, the rate function is given by

$$I(z) = \sup_{t \in \mathbb{R}} [zt - \log e^{\mu t + \frac{\sigma^2 t^2}{2}}] = \sup_{t \in \mathbb{R}} [zt - \mu t - \frac{\sigma^2 t^2}{2}] \ ,$$

Now, setting the derivative of this expression to zero we find that

$$\frac{\partial}{\partial t}(zt - \mu t - \frac{\sigma^2 t^2}{2}) = z - \mu - \sigma^2 t = 0 \ ,$$

which yields the solution $\tau = \frac{z-\mu}{\sigma^2}$. Thus, the rate function for a series of iid normally distributed RVs is

$$\begin{aligned} I(z) &= z\frac{z-\mu}{\sigma^2} - \mu\frac{z-\mu}{\sigma^2} - \frac{\sigma^2\frac{z-\mu}{\sigma^2}}{2} \\ &= \frac{(\sigma^2+2)(z-\mu)^2}{2\sigma^2} \ . \end{aligned}$$

and we can easily verify that $\tau = z$, so the rate function for a series of iid random variables with Z distributions is $I(z) = z^2/2$. In Fig 6, we see that the rate function is parabolic around the expected value, $\mu = 0$.

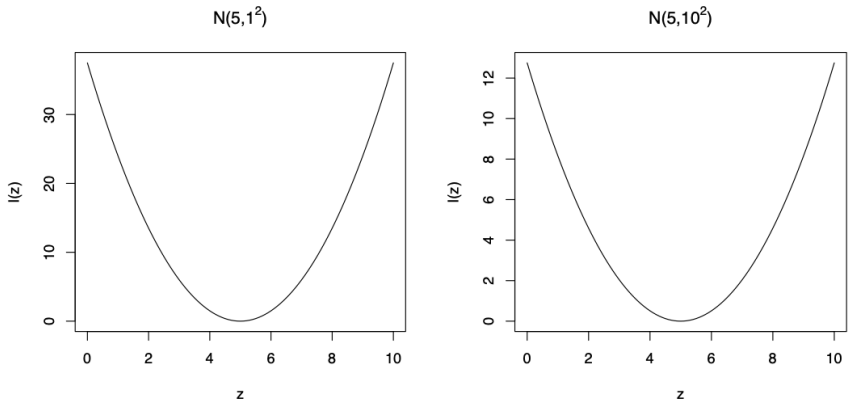

FIG. 6. The rate function for a series of normal iid random variables.

Exponential Distribution:

Now, let (X_n) be iid random variables such that $X_1 \sim \text{Expo}(\lambda)$. We showed earlier that $\varphi(t) = \frac{\lambda}{\lambda - t}$, and is defined for $t < \lambda$. Evoking the generalization stated in (27), we can see that $\varphi(0 \pm \epsilon) < \infty$, so $0 \in \mathcal{D}_\varphi$, satisfying the initial condition for Cramér's Theorem. Thus, the rate function is given by

$$I(z) = \sup_{t\in\mathbb{R}}[zt - \log\frac{\lambda}{\lambda - t}] = \sup_{t\in\mathbb{R}}[zt - \log\lambda + \log(\lambda - t)] \ .$$

Now, setting the derivative of this expression to zero we find that

$$\frac{\partial}{\partial t}(zt - \log\lambda + \log(\lambda - t)) = z - \frac{1}{\lambda - t} = 0 \ ,$$

which yields the unique solution $\tau = \lambda - 1/z$. Thus, the rate function for a series of iid RVs with Exponential distributions is

$$I(z) = z\lambda - 1 - \log\lambda - \log z \ .$$

In Fig 7, we see that the rate function has a minimum at $\frac{1}{\lambda}$, and has an asymptotic slope for $z \gg \lambda$.

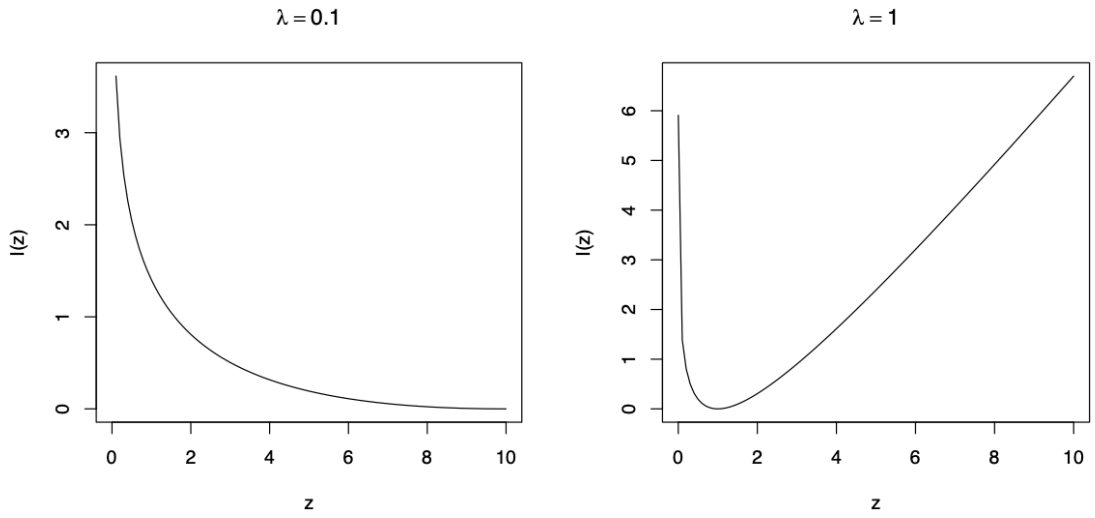

FIG. 7. The rate function for a series of exponential iid random variables with $\lambda = 0.1$ and $\lambda = 1$.

IV. CONCLUSION

We have laid out an introductory framework for large deviations. We have explored what large deviations are, and what they look like in different distributions of iid random variables. We found that the probabilistic decay rate for large deviations had exponential behavior across many distributions. Finally, we proved Cramér's Theorem, showing that the rate function for a series of iid random variables is the Fenchel-Legendre transform of the logarithm of the moment-generating function. Through learning about LDT, we extend our understanding of distributions of iid random variables beyond their central tendencies. LDT finds application in a variety of technical fields, including entropy in statistical mechanics, finance, and informational systems.

References

[1] Billingsley, P. 1979, *Probability and Measure*, The University of Chicago

[2] Grinstead, C., Snell, J. 2006, *Introduction to Probability*, Swarthmore College

[3] den Hollander, F. 2000, *Large Deviations*, American Mathematical Society

[4] Lasserre, J. B. 2009, *Linear and Integer Programming vs Linear Integration and Counting*, University of Toulouse

[5] Lewis, J. T., Russell, R. 1997, *An Introduction to Large Deviations for Teletraffic Engineers*, University of Warwick

[6] author not stated. 2013, *Lecture 2: Large Deviations for iid Random Variables*, lecture notes, Massachusetts Institute of Technology, delivered 9 September 2013