# Large Deviation Principle for i.i.d Random Variables

Leonardo T. Rolla

June 7, 2020

**Abstract**

Let $(X_n)_n$ be i.i.d. random variables and $S_n = X_1 + \cdots + X_n$. We state and prove Chernoff's Concentration Inequality for $\frac{S_n}{n}$ around $\mu$. We state Cramér's Large Deviation Principle for sums of

$$e^{-I(J^\circ)\cdot n + o(n)} \leqslant P\left(\tfrac{S_n}{n} \in J\right) \leqslant e^{-I(J^*)\cdot n + o(n)}, \quad J \text{ interval.}$$

We prove the lower bound assuming for simplicity that the supremum

$$I(a) = \sup_{t\in\mathbb{R}} \big[at - \log M(t)\big]$$

is a maximum. We prove the upper bound assuming for simplicity that $X$ has exponential moments.

Requirements:
– One-sided derivatives of moment generating functions
– Biasing the distribution of a random variable by an integrable function
– Hölder's inequality (for the accessory remark that $\log M$ is convex)

Notation. The term "$o(b_n)$" denotes a function $g(n)$ satisfying $\frac{g(n)}{b_n} \to 0$. This function depends on $a$, $\delta$, $J$ and on the distribution of $X$. Each time it appears, it denotes a different function.

# 1 Concentration Inequality

Let $X$ be a random variable with $EX = \mu$.

Let $M(t) = Ee^{tX}$, which is finite on an interval from $\mathcal{D}_M^- \leqslant 0$ to $\mathcal{D}_M^+ \geqslant 0$.

Let $X_n$ be i.i.d. and $S_n = X_1 + \cdots + X_n$.

The weak law of large number was proved as

$$P\left(\tfrac{S_n}{n} \geqslant a\right) \leqslant P\left[(\tfrac{S_n}{n} - \mu)^2 \geqslant (a-\mu)^2\right] \leqslant \tfrac{1}{(a-\mu)^2} E\left(\tfrac{S_n}{n} - \mu\right)^2 = \tfrac{VX}{(a-\mu)^2} \cdot \frac{1}{n}$$

So this means that when $EX^2 < \infty$ the probability of $\frac{S_n}{n}$ deviating from $\mu$ by a fixed amount $a - \mu$ decays at least as fast as $\frac{1}{n}$. In the proof of the LLN by Cantelli we see that when $EX^4 < \infty$ this probability decays at least as fast as $\frac{1}{n^2}$, and in general if $EX^{2k} < \infty$ it decays at least as fast as $\frac{1}{n^k}$.

We now try to do better using moments of $e^{tX}$ rather than $X^{2k}$. For $t \geqslant 0$,

$$P\left(\tfrac{S_n}{n} \geqslant a\right) \leqslant P\left[e^{tS_n} \geqslant e^{tan}\right] \leqslant \tfrac{1}{e^{tan}} Ee^{tS_n} = e^{-tan} M(t)^n = e^{-[at - \log M(t)]n}.$$

Likewise, for $a < \mu$ and $t \leqslant 0$,

$$P\left(\tfrac{S_n}{n} \leqslant a\right) \leqslant e^{-[at - \log M(t)]n}. \tag{1}$$

So if the expression in brackets can be made positive, we will have established that this probability in fact decays at least as fast as exponential.

**Theorem 2** (Chernoff's Concentration Inequality)**.** *If $\mathcal{D}_M^+ > 0$, then for any $a > \mu$ there is $t > 0$ such that $[at - \log M(t)] > 0$. In particular, $P\left(\frac{S_n}{n} \geqslant a\right)$ decays at least exponentially fast in $n$. Analogously for $a < \mu$ if $\mathcal{D}_M^- < 0$.*

*Proof.* Suppose $\mathcal{D}_M^+ > 0$ and let $a > \mu$. Then, as right-side derivatives,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[at - \log M(t)\right]\Big|_{t=0} = a - \frac{M'(0)}{M(0)} = a - \mu > 0,$$

so for $t$ positive and small the above expression is positive.

Similarly, if we suppose that $\mathcal{D}_M^- < 0$ and take $a < \mu$, then by taking left-side derivative we see that the term in bracket will be positive for $t$ negative and small. □

# 2 Large Deviation Principle

We start with the fundamental concept of rate function.

**Definition 3** (rate function)**.** Let $X$ be a random variable. We define the function $I$, the *rate function* associated to the distribution of $X$, by

$$I(a) = \sup_{t \in \mathbb{R}} \left[ at - \log M(t) \right].$$

We can think of the rate function as an attempt to make the most out of estimate (1). The reason why the function $I$ deserves this name is that, once we maximize $[at - \log M(t)]$ over all $t$, inequality (1) is not "yet another" upper bound, but actually the best possible estimate. The next theorem makes this claim precise. Given $A \subseteq \mathbb{R}$, to describe the "easiest way" for $\frac{S_n}{n}$ to be in $A$ we denote $I(A) = \inf_{a \in A} I(a)$.

**Theorem 4** (Cramér's Large Deviation Principle)**.** *Let $J$ be an interval. Denote by $J^\circ$ and $J^*$ the corresponding open and closed intervals. Then*

$$e^{-I(J^\circ)\cdot n + o(n)} \leqslant P\left(\tfrac{S_n}{n} \in J\right) \leqslant e^{-I(J^*)\cdot n + o(n)}.$$

*In particular, when $I(J^\circ) = I(J^*)$, we have the exact rate of exponential decay for these probabilities:*

$$P\left(\tfrac{S_n}{n} \in J\right) = e^{-I(J)\cdot n + o(n)}.$$

Before proving the above theorem, let us discuss the relation between $M$ and $I$, and its geometric interpretation.

**Proposition 5.** *The functions $I$ and $\log M$ are convex.*

*Proof.* We start with the convexity of $I$. Let $0 < \alpha < 1$ and $\beta = 1 - \alpha$. For $a_1$ and $a_2 \in \mathbb{R}$,

$$\begin{aligned} I(\alpha a_1 + \beta a_2) &= \sup_{t\in\mathbb{R}} \big[(\alpha a_1 + \beta a_2)t - (\alpha_1 + \alpha_2)M(t)\big] \\ &= \sup_{t\in\mathbb{R}} \big[\alpha(a_1 t - M(t)) + \beta(a_2 t - M(t))\big] \\ &\leqslant \sup_{t\in\mathbb{R}} \big[\alpha(a_1 t - M(t))\big] + \sup_{t\in\mathbb{R}} \big[\beta(a_2 t - M(t))\big] \\ &= \alpha I(a_1) + \beta I(a_2). \end{aligned}$$

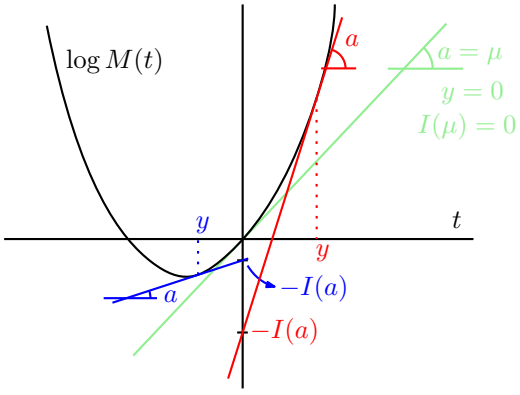
We now turn to the convexity of $M$. Let $t_1$ and $t_2 \in \mathbb{R}$. Using Hölder's inequality we get

$$\begin{aligned} \log M(\alpha t_1 + \beta t_2) &= \log E\left[e^{\alpha t_1 X} \cdot e^{\beta t_2 X}\right] \\ &\leqslant \log \left\{ \left[E\left(e^{\alpha t_1 X}\right)^{\frac{1}{\alpha}}\right]^{\alpha} \left[E\left(e^{\beta t_2 X}\right)^{\frac{1}{\beta}}\right]^{\beta} \right\} \\ &= \alpha \log E e^{t_1 X} + \beta \log E e^{t_2 X} \\ &= \alpha \log M(t_1) + \beta \log M(t_2). \end{aligned}$$

□

In case the supremum in the definition of $I(a)$ is attained at $y \in \mathbb{R}$, we have $0 = \frac{\mathrm{d}}{\mathrm{d}y}\left[ay - \log M(y)\right] = a - \frac{M'(y)}{M(y)}$, so $a = \frac{\mathrm{d}}{\mathrm{d}y} \log M(y) = \frac{M'(y)}{M(y)}$, and solving $y$ for $a$ we can sometimes compute

$$I(a) = a \cdot y - \log M(y), \quad y = y(a).$$



Let us illustrate this with a few examples.

If $X \sim \text{Poisson}(\lambda)$, we have

$$\log M(t) = \lambda(e^t - 1),$$

so

$$a = [\log M(y)]' = \lambda e^y, \qquad y = \log \tfrac{a}{\lambda},$$

and

$$I(a) = ay - \log M(y) = a \log \tfrac{a}{\lambda} - a + \lambda.$$

In fact,

$$I(a) = \begin{cases} a \log \frac{a}{\lambda} - a + \lambda, & a \geqslant 0, \\ +\infty & a < 0. \end{cases}$$

If $X \sim \mathcal{N}(\mu, 1)$, we have

$$\log M(t) = \frac{t^2}{2} + t\mu,$$

so

$$a = [\log M(y)]' = y + \mu, \qquad y = a - \mu,$$

and

$$I(a) = a(a-\mu) - \left[\tfrac{(a-\mu)^2}{2} + \mu(a-\mu)\right] = \frac{(a-\mu)^2}{2}, \quad a \in \mathbb{R}.$$

If $X \sim \text{Exp}(1)$, we have

$$M(t) = \frac{1}{1-t}, \quad t < 1,$$

so

$$a = \frac{M'(y)}{M(y)} = \frac{(1-y)^{-2}}{(1-y)^{-1}} = \frac{1}{1-y}, \qquad y = 1 - \frac{1}{a},$$

and

$$I(a) = ay - \log M(y) = a(1 - \tfrac{1}{a}) \log M(y) = a - 1 - \log a.$$

In fact,

$$I(a) = \begin{cases} a - 1 - \log a, & a > 0, \\ +\infty & a \leqslant 0. \end{cases}$$

If $X \sim \text{Bernoulli}(p)$, we have

$$M(t) = pe^t + 1 - p,$$

so

$$a = \frac{M'(y)}{M(y)} = \frac{pe^y}{pe^y + 1 - p}, \qquad y = \log(\tfrac{a}{p} \cdot \tfrac{1-p}{1-a}),$$

and

$$I(a) = ay - \log M(y) = \cdots = a \log \tfrac{a}{p} + (1-a) \log \tfrac{1-a}{1-p}.$$

In fact,

$$I(a) = \begin{cases} a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}, & 0 < a < 1, \\ \log \frac{1}{p}, & a = 1, \\ \log \frac{1}{1-p}, & a = 0, \\ +\infty & a < 0 \text{ or } a > 1. \end{cases}$$

If $X = \mu$ is constant, we have

$$\log M(t) = \mu t,$$

so

$$a = [\log M(t)]' = \mu, \qquad y \text{ can be any number},$$

and

$$I(a) = ay - \log M(y) = 0.$$

In fact,

$$I(a) = \begin{cases} 0, & a = \mu \\ +\infty & a \neq \mu. \end{cases}$$

# 3 Proof of lower bound

**Theorem 6.** *For any $a \in \mathbb{R}$ and $\delta > 0$,*

$$P\left(\tfrac{S_n}{n} \in [a-\delta, a+\delta]\right) \geqslant e^{-I(a)\cdot n + o(n)}.$$

The theorem is true as stated, with no assumptions on the distribution of $X$. However, we will assume that the supremum in $I(a)$ is attained, that is,

$$I(a) = a \cdot y - \log M(y)$$

for some $y \in (\mathcal{D}_M^-, \mathcal{D}_M^+)$. Dropping this assumption requires complicated technical steps that we cannot go into.

*Proof.* Since the supremum is attained in the interior of $(\mathcal{D}_M^-, \mathcal{D}_M^+)$, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\Big[at - \log M(t)\Big]\Big|_{t=y} = a - \frac{M'(y)}{M(y)},$$

and therefore

$$\frac{E[Xe^{yX}]}{E[e^{yX}]} = a.$$

The main observation is that the expression on the left-hand side corresponds to the expectation of a random variable $Y$ which is distributed as the variable $X$ biased by the factor $f(x) = e^{yx}$, $x \in \mathbb{R}$. That is, for a random variable $Y$ whose distribution is given by

$$P(Y \in B) = \tfrac{E[\mathbb{1}_B(X)e^{yX}]}{E[e^{yX}]} = E\left[\mathbb{1}_B(X)\tfrac{e^{yX}}{M(y)}\right],$$

we have $EY = a$. So for i.i.d. $Y_1, Y_2, \ldots$ distributed as this biased version of $X$, the occurrence of $\frac{Y_1+\cdots+Y_n}{n} \approx a$ is not a rare event.

The proof then consists in controlling the likelihood ratio of $(X_1, \ldots, X_n)$ with respect to $(Y_1, \ldots, Y_n)$ on a subset of $\mathbb{R}^n$ that is typical for the latter vector, so that such ratio does not get lower than $e^{-I(a)\cdot n - o(n)}$.

Fix an $\varepsilon \in (0, \delta]$, and define the set

$$B_n^\varepsilon = \left\{(z_1, \ldots, z_n) : \left|\tfrac{z_1+\cdots+z_n}{n} - a\right| \leqslant \varepsilon\right\} \subseteq \mathbb{R}^d.$$

Then

$$\begin{aligned}
P\left(\tfrac{S_n}{n} \in [a-\varepsilon, a+\varepsilon]\right) &= E\left[\mathbb{1}_{B_n^\varepsilon}(X_1, \ldots, X_n)\right] \\
&= E\left[\tfrac{M(y)^n}{e^{y(X_1+\cdots+X_n)}}\, \mathbb{1}_{B_n^\varepsilon}(X_1, \ldots, X_n) \tfrac{e^{yX_1}}{M(y)} \cdots \tfrac{e^{yX_n}}{M(y)}\right] \\
&\geqslant E\left[\frac{M(y)^n}{e^{ayn+|y|\varepsilon n}} \mathbb{1}_{B_n^\varepsilon}(X_1, \ldots, X_n) \tfrac{e^{yX_1}}{M(y)} \cdots \tfrac{e^{yX_n}}{M(y)}\right] \\
&= e^{-[ay - \log M(y) - |y|\varepsilon]\cdot n} \cdot P\left((Y_1, \ldots, Y_n) \in B_n^\varepsilon\right) \\
&= e^{-[ay - \log M(y) - |y|\varepsilon]\cdot n} \cdot P\left(\tfrac{Y_1+\cdots+Y_n}{n} \in [a-\varepsilon, a+\varepsilon]\right).
\end{aligned}$$

The latter probability converges to 1 by the Law of Large Numbers, thus

$$P\left(\left|\tfrac{S_n}{n} - a\right| \leqslant \delta\right) \geqslant P\left(\left|\tfrac{S_n}{n} - a\right| \leqslant \varepsilon\right) \geqslant e^{-I(a)\cdot n - 2|y|\varepsilon\cdot n}$$

for large enough $n$. Since $\varepsilon \in (0, \delta]$ was arbitrary, this gives

$$P\left(\left|\tfrac{S_n}{n} - a\right| \leqslant \delta\right) \geqslant e^{-I(a)\cdot n + o(n)},$$

completing the proof. □

*Proof of the lower bound in Theorem 4.* Let $\varepsilon > 0$. Take $a \in J^\circ$ such that $I(a) \leqslant I(J^\circ) + \varepsilon$. Take $\delta > 0$ such that $[a-\delta, a+\delta] \subseteq J$. Then using Theorem 6 we get

$$P\left(\tfrac{S_n}{n} \in J\right) \geqslant P\left(\tfrac{S_n}{n} \in [a-\delta, a+\delta]\right) \geqslant e^{-I(a)\cdot n + o(n)} \geqslant e^{-I(J^\circ)\cdot n - \varepsilon n + o(n)}.$$

Since $\varepsilon$ is arbitrary, $P\left(\tfrac{S_n}{n} \in J\right) \geqslant e^{-I(J^\circ)\cdot n + o(n)}$, completing the proof. □

# 4 Proof of upper bound

The upper bound in Theorem 4 is a direct consequence of Theorem 2, Jensen's inequality and convexity of $I$.

There are three cases depending on where $M$ is finite. The main case is $\mathcal{D}_M^- < 0 < \mathcal{D}_M^+$. The trivial case is $\mathcal{D}_M^- = 0 = \mathcal{D}_M^+$, which implies that $I$ is identically 0 so $P\left(\frac{S_n}{n} \in J\right) \leqslant e^{-I(J)\cdot n}$ always holds. The case $\mathcal{D}_M^- < 0 = \mathcal{D}_M^+$ or $\mathcal{D}_M^- = 0 < \mathcal{D}_M^+$ can have interesting behavior, and it has subcases depending on whether $\mu$ is finite.

The proof uses monotonicity properties of the rate function that follow from convexity. We consider the main case only. Proving the general result would require the study of convexity properties of $I$ for all corner cases.

**Proposition 7.** *Suppose that $\mathcal{D}_M^- < 0 < \mathcal{D}_M^+$. Then the rate function $I$ is increasing on $[\mu, +\infty)$ and decreasing on $(-\infty, \mu]$. Moreover, $I(\mu) = 0$ and*

$$I(a) = \sup_{t \geqslant 0}[at - \log M(t)] \text{ for } a \geqslant \mu$$

*and*

$$I(a) = \sup_{t \leqslant 0}[at - \log M(t)] \text{ for } a \leqslant \mu.$$

*Proof.* Taking $t = 0$ we have $[at - \log M(t)] = 0$, so $I(a) \geqslant 0$ for all $a$. Now, by Jensen's inequality, $M(t) = Ee^{tX} \geqslant e^{EtX} = e^{t\mu}$, so

$$\mu t - \log M(t) \leqslant 0.$$

This implies that $I(\mu) = 0$. It also implies that, for $a > \mu$ and $t < 0$, $at - \log M(t) < 0$, so $I(a) = \sup_{a \geqslant 0}[at - \log M(t)]$. Analogously, for $a < \mu$ we have $I(a) = \sup_{a \leqslant 0}[at - \log M(t)]$.

By Theorem 2, $I(y) > 0$ for all $y \neq \mu$. In particular, for any $y > x > \mu$, we have $I(\mu) = 0 < I(x) \leqslant \frac{x-\mu}{y-\mu} I(y) < I(y)$, and therefore $I$ is increasing on $[\mu, \infty)$. Analogously, if $y < x < \mu$, we have $I(\mu) = 0 < I(x) \leqslant \frac{\mu - x}{\mu - y} I(y) < I(y)$, so $I$ is decreasing on $(-\infty, \mu]$. □

*Proof of the upper bound in Theorem 4.* Write $J^* = [c, a] \subseteq \mathbb{R}$. If $c \leqslant \mu \leqslant a$, $I(J^*) = 0$ and there is nothing to prove. So we can assume that $a < \mu$, as the case $c > \mu$ is handled similarly. Let $\varepsilon > 0$. By Proposition 7, we have

$$I(J^*) = I(a) = \sup_{t \geqslant 0}[at - \log M(t)],$$

so we can take $t \geqslant 0$ such that $[at - \log M(t)] \geqslant I(J^*) - \varepsilon$. Now using estimate (1) we get

$$P\left(\tfrac{S_n}{n} \in J\right) \leqslant P\left(\tfrac{S_n}{n} \leqslant a\right) \leqslant e^{-I(J^*)\cdot n - \varepsilon n}.$$

Since $\varepsilon$ is arbitrary, $P\left(\tfrac{S_n}{n} \in J\right) \leqslant e^{-I(J^*)\cdot n + o(n)}$, completing the proof. $\square$

# Hölder's inequality

Let $p \geqslant 1$, $q \geqslant 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$.

**Proposition 8** (Young's inequality)**.** *For $a, b \geqslant 0$, $ab \leqslant \dfrac{a^p}{p} + \dfrac{b^q}{q}$*

**Proposition 9** (Hölder's inequality)**.** *If $X$ and $Y$ have finite $p$- and $q$-moments, respectively, then $XY$ is integrable and $E[XY] \leqslant \|X\|_p \cdot \|Y\|_q$.*

# References


[DZ10] A. Dembo, O. Zeitouni. *Large deviations techniques and applications*, vol. 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.

[Gam13] D. Gamarnik. *15.070j advanced stochastic processes*, 2013. MIT OpenCourseWare.

[OV05] E. Olivieri, M. E. Vares. *Large deviations and metastability*, vol. 100 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2005.