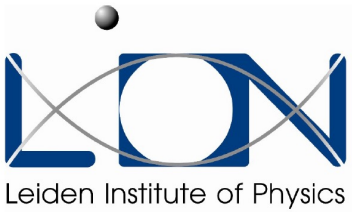
An introduction to large deviation theory - I

Jinfu Chen, June 19 @Thursday Lecture

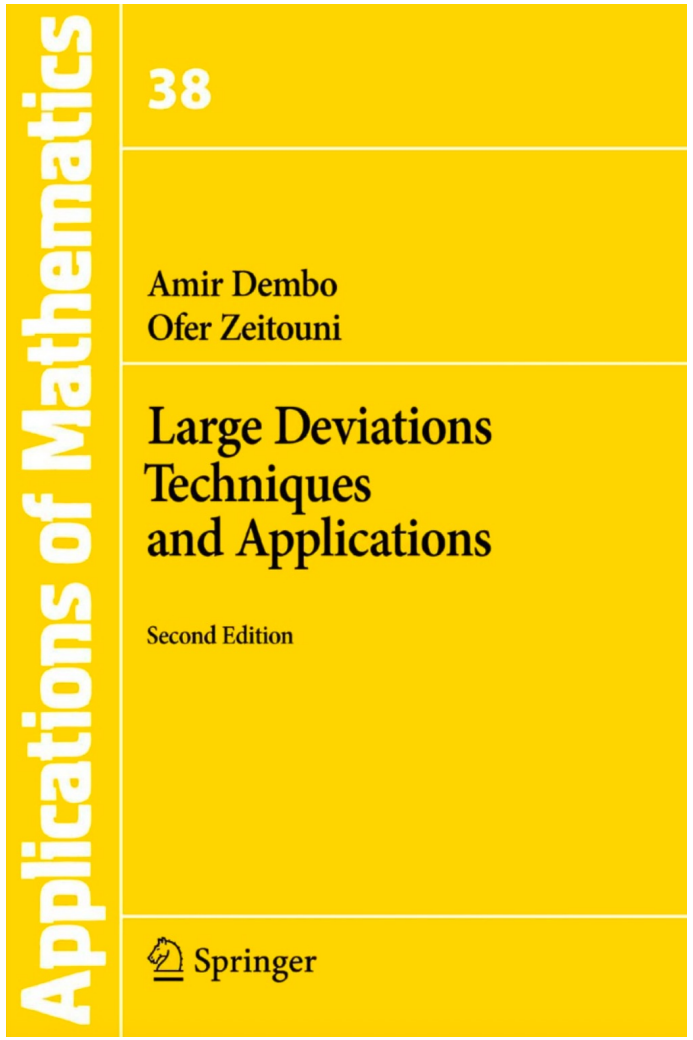


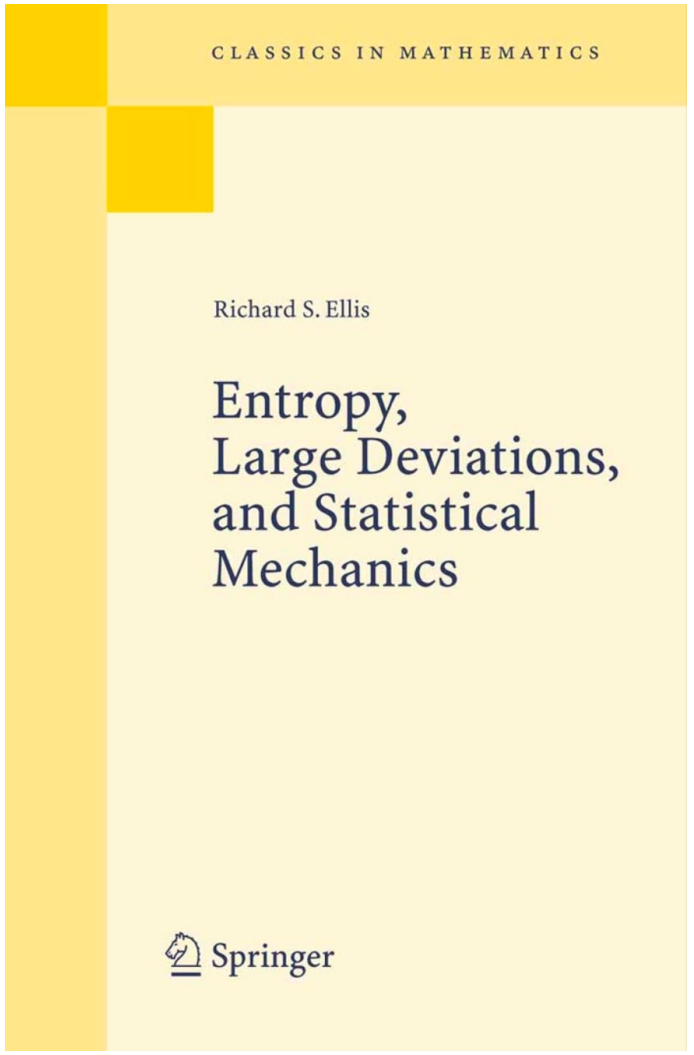
Universiteit Leiden
The Netherlands





References for Large deviation Theory





References for Large deviation Theory

- H. Touchette, The large deviation approach to statistical mechanics, Phys. Rep. 478, 1 (2009).

Physics Reports 478 (2009) 1–69

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep



The large deviation approach to statistical mechanics

Hugo Touchette *

School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

Motivation & Overview

- Whenever physicists **calculate an entropy or free energy function**, large deviation theory is at play.

- Large deviation theory provides a rigorous mathematical description for **statistical physics**, for example, to better understand the maximum entropy principle.

- Large deviation theory is also involved when studying the properties of **many-particle systems**, whether in equilibrium or not.

Motivation & Overview

- For mathematicians, the large deviation is

  - A theory describing the **exponential** decay of probabilities of rare.

  - A calculus of exponential-order probabilities based on **saddle-point approximations**.

  - A generalization of **Cramér's Theorem** for sample means of i.i.d. random variables.

  - An extension or refinement of the **Law of Large Numbers** and **the Central Limit Theorem**.

Motivation & Overview

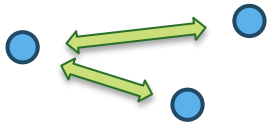
- For physicists, the large deviation is

  - A collection of techniques for calculating entropies and free energies.

  - A rigorous version of the **saddle-point approximation** (large n limit).

  - A rigorous formulation of **statistical mechanics**.

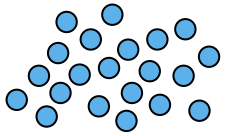
Microscopic models



Thermodynamic limit ($n \to \infty$)

Large deviation theory

Macroscopic properties



Content

- A fast review of probability theory
  - **Law of Large Numbers** and **Central Limit Theorem**

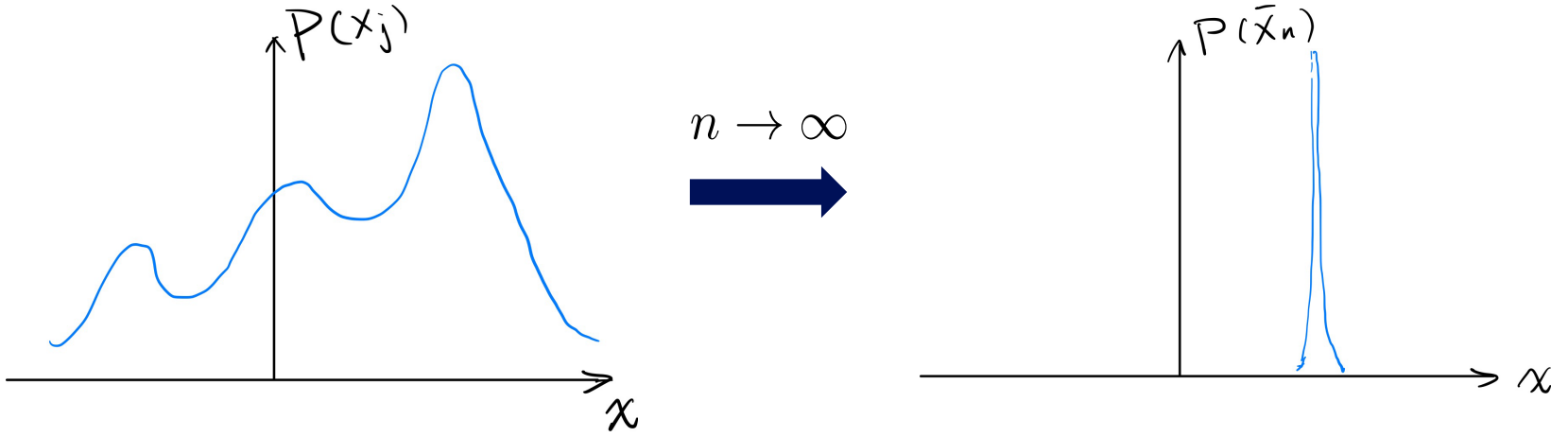
- Large Deviation Principle
  - **Gärtner–Ellis Theorem** – Method for evaluating rate functions
  - **Cramér's Theorem** – Level-1 of large deviations
  - Properties of scaled cumulant generating function $\lambda(k)$ and rate function $I(a)$

- **Sanov's Theorem** – Level-2 large deviations
- Take-home messages

Law of Large Numbers (LLN):

- independent identically distributed (**i.i.d.**) random variables X_j
- Sample mean $\bar{X}_n = \sum_{j=1}^{n} X_j \,/n$ converges to expectation μ with high probability.

Weak LLN	Strong LLN
$\bar{X}_n \xrightarrow{P} \mu, \text{ when } n \to \infty$	$\bar{X}_n \xrightarrow{\text{a.s.}} \mu, \text{ when } n \to \infty$
$\lim_{n\to\infty} \Pr(\lvert\bar{X}_n - \mu\rvert < \varepsilon) = 1$	$\Pr(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$



Central Limit Theorem:

- The normalized sum converges **in distribution** to a normal distribution as $n \to \infty$

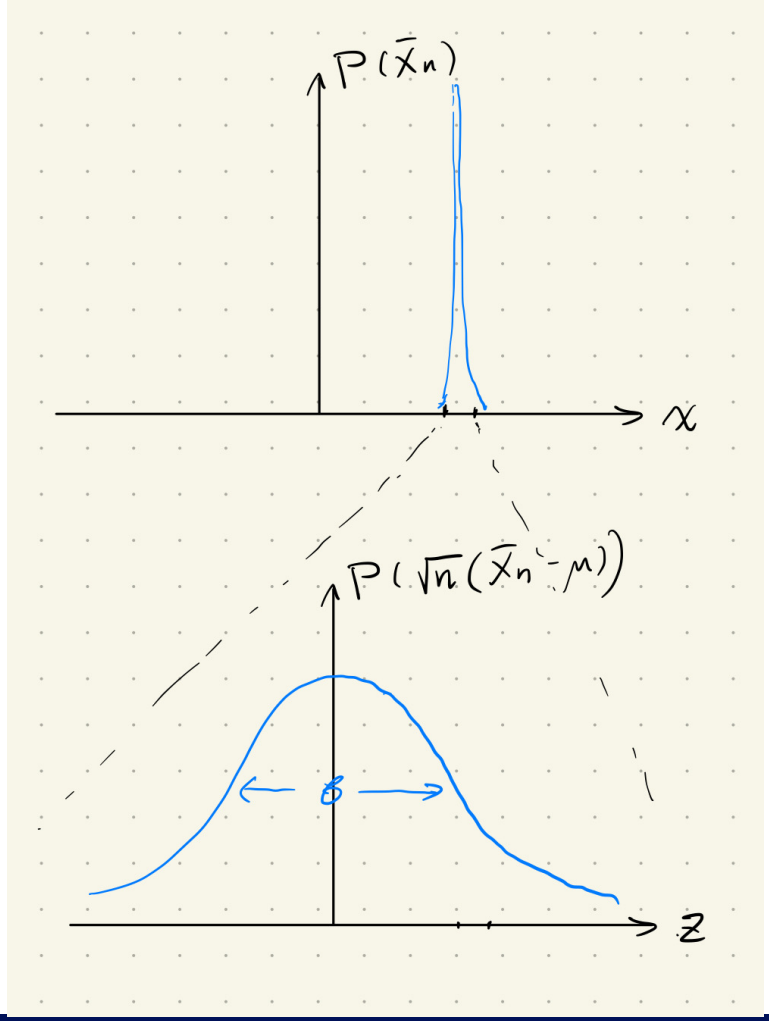
Central Limit Theorem

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

$$\lim_{n\to\infty} \Pr\left[\sqrt{n}\left(\bar{X}_n - \mu\right) \leq z\right]$$

$$= \lim_{n\to\infty} \Pr\left[\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leq \frac{z}{\sigma}\right]$$

$$= \Phi\left(\frac{z}{\sigma}\right)$$

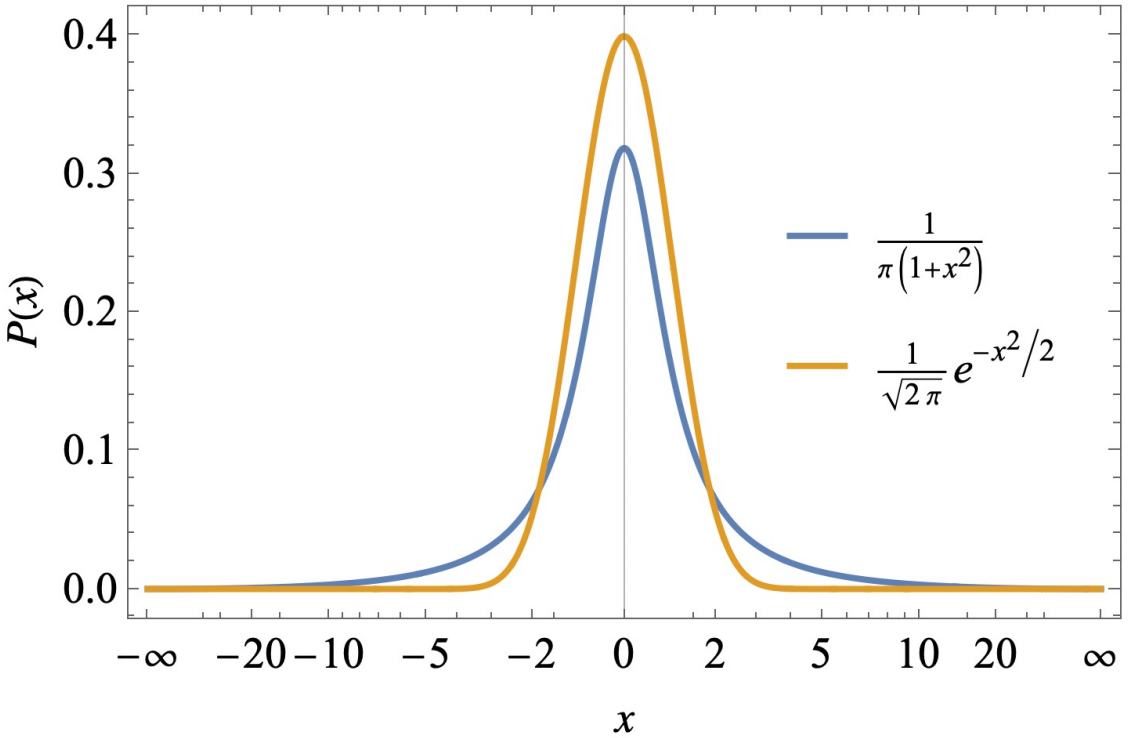


Remark

- The validity of the Law of Large Numbers and the Central Limit Theorem requires the existence of some moments of the random variables.
- Existence of variance ➡ weak LLN and Central Limit Theorem

↘ + i.i.d. ➡ strong LLN

- Counterexample: **Cauchy distribution**



- To make life easier, we focus on light-tailed distributions in the following.

What is Large Deviation Principle?

- **Level-1 Large deviation principle:**
- How probability of the sample mean $\bar{X}_n = \sum_{j=1}^{n} X_j \,/n$ concentrates around μ

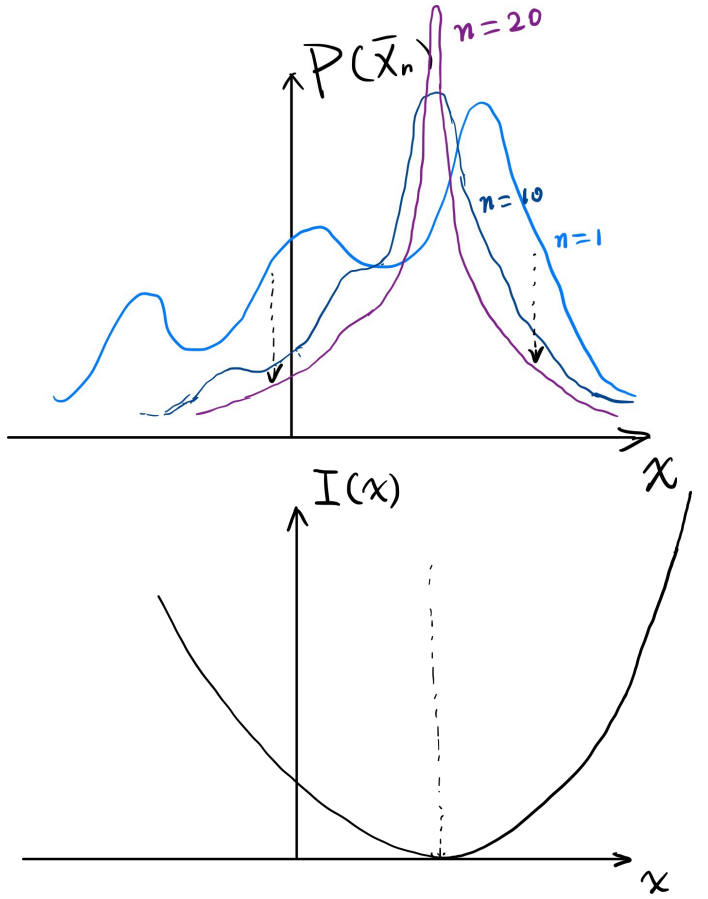
$$I(x) \equiv -\lim_{n\to\infty} \frac{1}{n} \ln\big(P(\bar{X}_n = x)\big)$$

- With increasing n, the probability concentrates as

$$P(\bar{X}_n = x) \propto e^{-nI(x)}$$

- More rigorously,

$$\lim_{n\to\infty} \frac{1}{n} \ln\big(P(\bar{X}_n \in A)\big) = -\inf_{x\in A} I(x)$$



- I will introduce **level-1 (Cramér's Theorem)** and **level-2 (Sanov's Theorem)** large deviation principles, and method **(Gärtner–Ellis Theorem)** to evaluate the rate functions.

- We focus on **i.i.d. random variables**. The results are also applicable to broader contexts in statistical physics.

- We will evaluate with simple examples:
- Gaussian sample mean, Random bits, Symbol frequencies, Exponential sample mean.

Gaussian Sample Mean

- Sequence: $X = (X_1, X_2, \cdots, X_n)$
- i.i.d. Gaussian random variables $X_i \sim \mathcal{N}(\mu, \sigma^2)$

$$p(X_i = x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

- The sample mean $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is also Gaussian
- Its distribution $p(S_n = s) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left[-n\frac{(s-\mu)^2}{2\sigma^2}\right] \approx e^{-nI(s)}$
- Rate function: $I(s) = \frac{(s-\mu)^2}{2\sigma^2}$

Random bits

- Sequence: $b = (b_1, b_2, \cdots, b_n)$
- i.i.d. random bits $b_j \sim \text{Bernoulli}(p)$

$$P(b_j = 1) = p,\ P(b_j = 0) = 1 - p$$

- The sample mean $R_n = \frac{1}{n}\sum_{i=1}^{n} b_i$

$$P(R_n = r) = p^{rn}(1-p)^{(1-r)n}\frac{n!}{(rn)!(n-rn)!}$$

- Law of Large Number: $R_n \overset{\text{a.s.}}{\to} p \quad \text{when } n \to \infty$

- Central Limit Theorem: $\sqrt{n}(R_n - p) \overset{d}{\to} \mathcal{N}(0, p(1-p))$

Random bits – Large Deviation

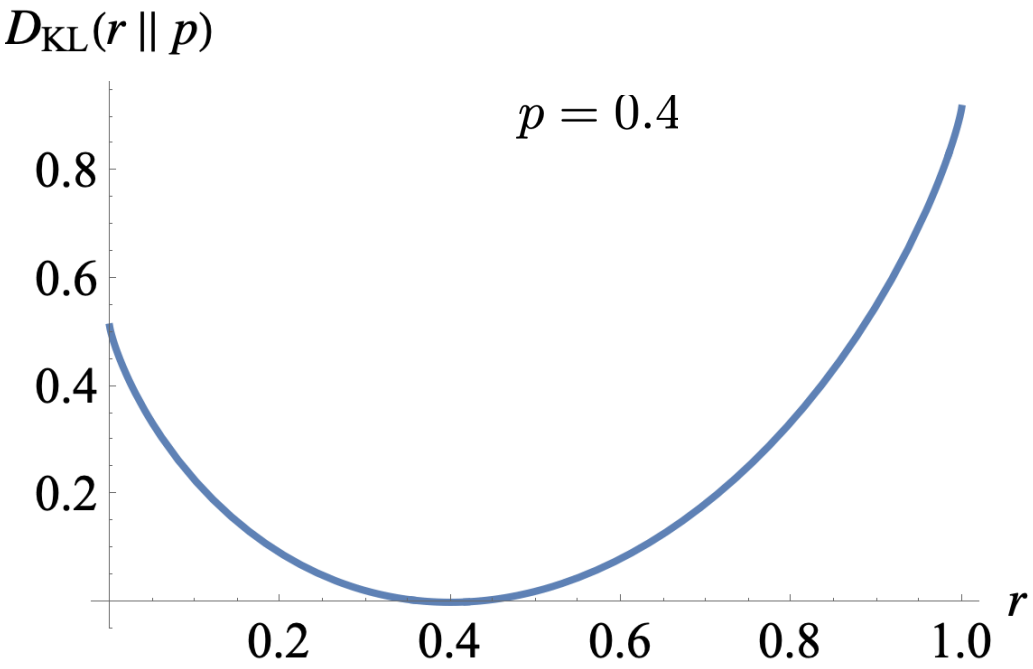
- We use the **Strling formula**

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

$$\begin{aligned}
\frac{n!}{(rn)!(n-rn)!} &\approx \frac{n^n e^{-n}\sqrt{2\pi n}}{(rn)^{rn} e^{-rn}\sqrt{2\pi rn}(n-rn)^{n-rn} e^{-n+rn}\sqrt{2\pi(n-rn)}} \\
&= \frac{e^{n\ln n - rn\ln(rn) - (n-rn)\ln(n-rn)}}{\sqrt{2\pi n r(1-r)}} \\
&= \frac{e^{-n[r\ln(r)+(1-r)\ln(1-r)]}}{\sqrt{2\pi n r(1-r)}}
\end{aligned}$$

Random bits – Large Deviation

$$
\begin{aligned}
& P(R_n = r) \\
=& p^{rn}(1-p)^{(1-r)n}\frac{n!}{(rn)!(n-rn)!} \\
=& \frac{e^{-n[r\ln(r)+(1-r)\ln(1-r)-r\ln p-(1-r)\ln(1-p)]}}{\sqrt{2\pi nr(1-r)}} \\
=& \frac{e^{-nD_{\mathrm{KL}}(r||p)}}{\sqrt{2\pi nr(1-r)}}
\end{aligned}
$$



- The rate function is

$$
\begin{aligned}
I(r) &= D_{\mathrm{KL}}(r||p) \\
&= r\ln\frac{r}{p} + (1-r)\ln\frac{1-r}{1-p}
\end{aligned}
$$

Large deviation principle

- Let A_n be a random variable indexed by the integer n. $P(A_n \in B)$ satisfies a large deviation principle with rate $I(B)$ if the limit exist

$$\lim_{n\to\infty} -\frac{1}{n} \ln P\left(A_n \in B\right) = I(B)$$

- We use the following notation

$$P(A_n \in \mathrm{d}a) \asymp e^{-nI(a)}\mathrm{d}a$$

$$a_n \asymp b_n \iff \lim_{n\to\infty} \frac{1}{n} \ln a_n = \lim_{n\to\infty} \frac{1}{n} \ln b_n$$

Some remarks

- Non-existence of limit:
- (1) Singular behavior
- (2) Super-exponential in n, $I = \infty$, e.g., e^{-Cn^2} or $e^{-Cn \ln n}$
- (3) Sub-exponential in n, $I = 0$, e.g., $e^{-C \ln n}$ (heavy-tail distribution)

- In some cases, the limit may not exist, but upper and lower bounds can be found

$$\mathrm{e}^{-nI_B^-} \leq P\left(A_n \in B\right) \leq \mathrm{e}^{-nI_B^+}.$$

- Discrete random variables become continuous in the limit $n \to \infty$.
- Instead of system size, n can also be other quantities.

Calculating rate functions

- **The large deviation theory addresses two main goals:**

- (1) Existence of a large deviation principle for random variables.

- (2) Derivation to the corresponding rate function.

- **For non-i.i.d. cases,** a more general approach is provided by the **Gärtner–Ellis Theorem.** We introduce in the following.

Gärtner–Ellis Theorem

- For a real random variable A_n, define the **scale cumulating generating function**

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle e^{nkA_n} \right\rangle$$

$$\left\langle e^{nkA_n} \right\rangle = \int_{\mathbb{R}} e^{nka} P(A_n \in \mathrm{d}a)$$

- **Gärtner–Ellis Theorem**: If $\lambda(k)$ exists and is differentiable for all $k \in \mathbb{R}$, then A_n satisfies a large deviation principle, i.e.,

$$P(A_n \in \mathrm{d}a) \asymp e^{-nI(a)} \mathrm{d}a$$

- The rate function is given by the Legendre–Fenchel transform (Gärtner–Ellis Theorem)

$$I(a) = \sup_{k\in\mathbb{R}} \{ka - \lambda(k)\}$$

An argument

- We verify the rate function under the assumption $P(A_n \in \mathrm{d}a) \asymp e^{-nI(a)}\mathrm{d}a$

- The generating function is

$$\left\langle e^{nkA_n} \right\rangle \asymp \int_{\mathbb{R}} e^{n[ka - I(a)]}\mathrm{d}a$$

$$\asymp \exp\left(n \sup_{a\in\mathbb{R}} \{ka - I(a)\} \right)$$

- The scaled cumulant generating function satisfies **Varadhan's Theorem**

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkA_n} \right\rangle = \sup_{a\in\mathbb{R}} \{ka - I(a)\}.$$

- Legendre–Fenchel transforms can be inverted when $\lambda(k)$ is differentiable

$$I(a) = \sup_{k\in\mathbb{R}} \{ka - \lambda(k)\}$$

Application: Cramér's Theorem

- **Cramér's Theorem**: For **i.i.d.** case, the generating function of the sample mean satisfies

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle e^{k\sum_{i=1}^{n} X_i} \right\rangle = \lim_{n\to\infty} \frac{1}{n} \ln \prod_{i=1}^{n} \left\langle e^{kX_i} \right\rangle = \ln \left\langle e^{kX} \right\rangle$$

- It is just the generating (characteristic) function of X_i
- **Gaussian**:

$$\lambda(k) = \ln \left\langle e^{kX} \right\rangle = \mu k + \frac{1}{2}\sigma^2 k^2$$

$$I(s) = \sup_k \{ks - \lambda(k)\} = \sup_k \{k(s-\mu) - \frac{1}{2}\sigma^2 k^2\} = \frac{(s-\mu)^2}{2\sigma^2}$$

Application: Cramér's Theorem

- **Random bits**:

$$\lambda(k) = \ln\left\langle e^{kb}\right\rangle = \ln[(1-p)e^{k\cdot 0} + pe^{k\cdot 1}]$$

$$\begin{aligned} I(r) &= \sup_k\{kr - \lambda(k)\} \\ &= \sup_k\{kr - \ln\left(1-p+pe^k\right)\} \\ &= r\ln\frac{r}{p} + (1-r)\ln\frac{1-r}{1-p} \end{aligned} \qquad k = \ln\frac{(1-p)r}{p(1-r)}$$

- The Gärtner–Ellis Theorem is not limited to i.i.d. random variables
- It can be applied to any random variable when $\lambda(k)$ exists and can be calculated.
- This is the situation in statistical physics for evaluating the partition function!

Properties of λ

- Derivatives

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle e^{nkA_n} \right\rangle$$

$$\lambda'(0) = \lim_{n\to\infty} \left. \frac{\left\langle A_n e^{nkA_n} \right\rangle}{\left\langle e^{nkA_n} \right\rangle} \right|_{k=0} = \lim_{n\to\infty} \langle A_n \rangle$$

$$\lambda''(0) = \lim_{n\to\infty} n(\left\langle A_n^2 \right\rangle - \langle A_n \rangle^2) = \lim_{n\to\infty} n\mathrm{var}(A_n)$$

- Convexity

$$\alpha \ln \left\langle \mathrm{e}^{nk_1 A_n} \right\rangle + (1-\alpha) \ln \left\langle \mathrm{e}^{nk_2 A_n} \right\rangle \geq \ln \left\langle \mathrm{e}^{n[\alpha k_1 + (1-\alpha)k_2] A_n} \right\rangle$$



$$\alpha\lambda(k_1) + (1-\alpha)\lambda(k_2) \geq \lambda(\alpha k_1 + (1-\alpha)k_2)$$

Properties of I

- Positivity: $I(a) \geq 0$

- **Law of Large Numbers** $\quad I(\mu) = 0 \quad \mu = \lim_{n\to\infty} \langle A_n \rangle$

- Let B be any set of A_n, then

$$P\left(A_n \in B\right) = \int_B P\left(A_n \in \mathrm{d}a\right) \asymp \int_B \mathrm{e}^{-nI(a)}\mathrm{d}a \asymp \mathrm{e}^{-n \inf_{a\in B} I(a)}$$

- $P(A_n \in B) \to 0$ exponentially fast with $n \to \infty$ if $\mu \notin B$

- **Central Limit Theorem** $\quad I''(a=\mu) = \dfrac{1}{\sigma^2}$

$$I(a) \approx \frac{1}{2} I''(a=\mu)(a-\mu)^2$$

$$P(A_n \in \mathrm{d}a) \approx e^{-nI''(a=\mu)(a-\mu)^2/2}\mathrm{d}a$$

A classic counterexample

- i.i.d. heavy-tail random variables under Pareto density

$$p(x) = \frac{A}{(|x| + c)^{\beta}}$$

$c > 0$
A normalized constant

- $\beta > 3$, the Central Limit Theorem holds for $n^{1/2} S_n$

- $I(a) = 0$ everywhere

- $\lambda(k)$ diverges for all $k \in \mathbb{R}$ except $k = 0$

Symbol Frequency

- i.i.d. discrete random variables $\omega = (\omega_1, \omega_2, \cdots, \omega_n)$
- Their values $\Lambda = \{1, 2, \cdots, q\}$ $P(\omega_i = j) = \rho_j > 0$
- The relative frequency with the symbol $j \in \Lambda$ is

$$L_{n,j}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i, j}$$

- The normalized vector: $L_n(\omega) = (L_{n,1}(\omega), L_{n,2}(\omega), \cdots, L_{n,q}(\omega))$
- Distribution of L_n: multinomial distribution

$$P(L_n = l) = \frac{n!}{\prod_{j=1}^{q} (nl_j)!} \prod_{j=1}^{q} \rho_j^{nl_j}$$

Sanov's Theorem – Level-2 large deviations

- According to results of the random bits, you can guess that

$$P(L_n = l) \approx e^{-nI_\rho(l)}, \quad I_\rho(l) = D_{KL}(l||\rho) = \sum_{j=1}^{q} l_j \ln \frac{l_j}{\rho_j}$$

- The rate function is related to the relative entropy or Kullback–Leibler distance
- The complete large deviation principle for L_n is known in large deviation theory as Sanov's Theorem.
- It describes the probability that the empirical distribution of i.i.d. samples deviates from the true distribution.

Contraction principle

- $I_A(a)$ is the rate function for A_n
- We want to find the rate function for $B_n = h(A_n)$. We call h a contraction of A_n

$$P(B_n \in \mathrm{d}b) = \int_{\{a:h(a)=b\}} P(A_n \in \mathrm{d}a)$$

$$e^{-nI_B(b)}\mathrm{d}b \asymp \exp\left(-n \inf_{a:h(a)=b} I_A(a)\right)\mathrm{d}a$$

$$I_B(b) = \inf_{a:h(a)=b} I_A(a)$$

Contraction from Level-2 to Level-1

$$\omega = (\omega_1, \omega_2, \cdots, \omega_n)$$

$$S_n = \frac{1}{n}\sum_{i=1}^{n} \omega_i = \sum_{j=1}^{q} j L_{n,j}(\omega)$$

- Its value $s = \sum_{j=1}^{q} j l_j$

- The contraction principle leads to

$$I_1(s) = \inf_{l:\sum_{j=1}^{q} j l_j = s} I_2(l)$$

Large deviation principle in 3 levels

- Introduced by Donsker and Varadhan

- Large deviations are studied at **three levels**:

  - **Level 1:** the sample mean $\bar{X}_n$ (Cramér's Theorem).

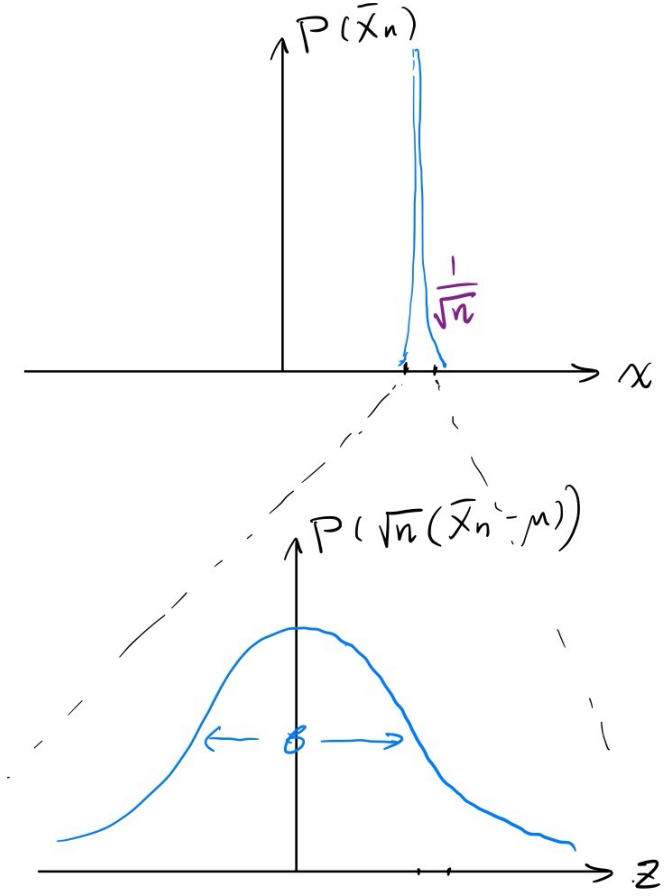
  - **Level 2:** the empirical vector L_n (Sanov's Theorem).

  - **Level 3:** the pair empirical distribution by including m-tuple empirical distribution.

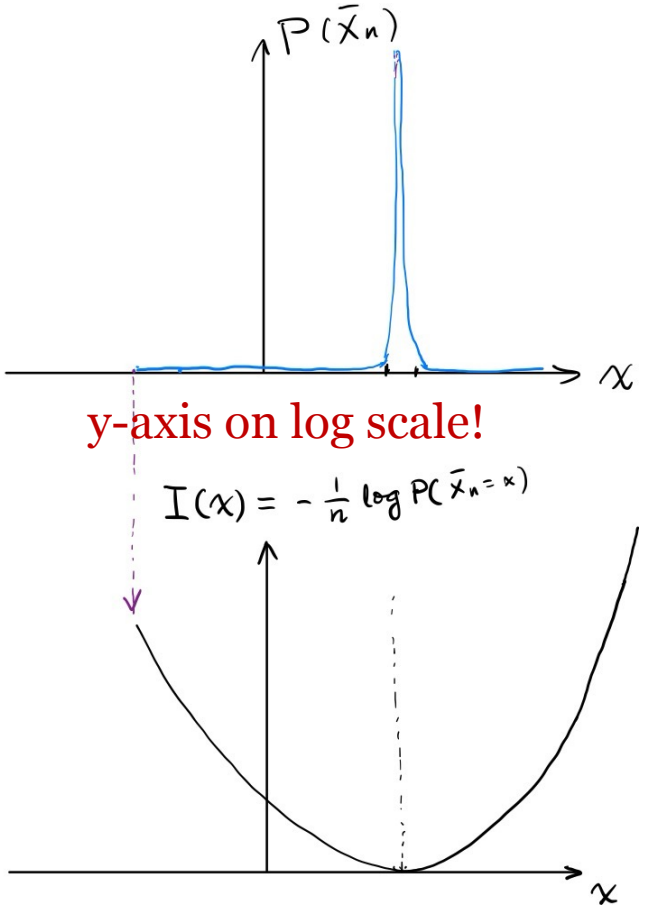
Summary

- In this lecture, we use only i.i.d. random variables for illustration of the large deviation theory.
- It can also be applied to more complex models, for example, Ising models, either fully connected (Curie Weiss) or finite dimension
- The large deviation theory can be used to study nonequilibrium processes, for example, relaxation processes [see J. Meibohm and M. Esposito, PRL 128, 110603 (2022)]
- Metastability is related to the non-convexity in the rate function (recall yesterday's talk).

Take-home message



Small deviation:
Central Limit Theorem



Large deviation:
Rate Function

Take-home message

Statistical physics	Large deviation theory
Entropy per site	minus Large deviation rate function
Free energy per site	Scale cumulating generating function
Most probable state	argmin $I(x)$
Second-order phase transition	Non-convex $I(x)$
First-order phase transition	Sudden change of argmin $I(x)$



Thank You

Universiteit Leiden
The Netherlands

Proof of Properties of I

- Positivity: $I(a) \geq 0$

$$\lambda(0) = \sup_a \{-I(a)\} = -\inf_a I(a) = -I(a^*) = 0$$

- Law of large numbers $\quad a^* = \lambda'(0) = \lim_{n\to\infty} \langle A_n \rangle$

$$\begin{aligned} I(a^*) &= \sup_k \{ka^* - \lambda(k)\} \\ &= k(a^*)a^* - \lambda(k(a^*)) \\ &= 0a^* - 0 \\ &= 0 \end{aligned}$$

$$k(a^*) = I'(a^*) = 0$$

$$\lambda(0) = \lim_{n\to\infty} \frac{1}{n} \ln \langle 1 \rangle = 0$$

- Let B be any set of A_n, then

$$P\left(A_n \in B\right) = \int_B P\left(A_n \in \mathrm{d}a\right) \asymp \int_B \mathrm{e}^{-nI(a)} \mathrm{d}a \asymp \mathrm{e}^{-n \inf_{a\in B} I(a)}$$

- $P(A_n \in B) \to 0$ exponentially fast with $n \to \infty$ if $a^* \notin B$

Next lecture

- Large deviations in equilibrium statistical mechanics
  - Entropy as a rate function
  - Free energy as a scaled cumulant generating function
  - Phase transitions
  - Legendre transforms in thermodynamics
  - Jaynes's Maximum Entropy Principle

$$dU\left(S,V,N_i\right) = TdS - pdV + \sum_i \mu_i dN_i$$

$$dF\left(T,V,N_i\right) = -SdT - pdV + \sum_i \mu_i dN_i$$

$$dH\left(S,p,N_i\right) = TdS + Vdp + \sum_i \mu_i dN_i$$

$$dG\left(T,p,N_i\right) = -SdT + Vdp + \sum_i \mu_i dN_i$$

- Markov processes, level 3 of large deviations

- Examples?