APMA2980: large deviation theory

Contents

Syllabus

The reading course will be an introduction to large deviation theory, using the following books/article as main reference:

- Large Deviations by Frank den Hollander [2],
- Analysis and Approximation of Rare Events by Amarjit Budhiraja and Paul Dupuis [1].,

in addition to several other papers, cited below accordingly.

Week #	Date	Topic	Reading
1	01/24 - 01/26	Theorem of Cramér and Sanov	dH Chapter 1 & 2
2	01/29 - 02/02	Sanov's Theorem via weak convergence	BD Chapter 3.1
3	02/05 - 02/09	General large deviation principle	dH Chapter 3
4	02/12 - 02/16	Large deviation for Markov Chains	dH Chapter 4
5	02/21 - 02/23	Gartner-Ellis Theorem & Hypothesis testing	dH Chapter 5 & 9
6	02/26 - 03/01	No meeting (D.C. travelling)	
7	03/04 - 03/08	Representation for functionals of Brownian motion	BD Chapter 3.2
8	03/11 - 03/15	Interacting diffusion	dH Chapter 10
9	03/18 - 03/22	Importance sampling for rare events	[3]
10	03/25 - 03/29	Spring Break!	
11	04/01 - 04/05	Importance sampling continued	
12	04/08 - 04/12	Cramer's theorem is atyipcal!	[4]
13	04/15 - 04/19	Representation for functionals of Poisson Processes	BD Chapter 3.3
14	04/22 - 04/26	LDP for finite-state Markov chains	BD Chapter 13.3
15	04/28 - 05/07	Weighted serve-the-longest-queue	BD Chapter BD Chapter 13.2

The reading group will meet from 11 a.m. to 12 p.m. on Wednesdays. Exercises and additional reading materials will be posted in the drop box folder.

1 Canonical Examples of Large Deviation

Remark 1.1 (to add...).

1. geometry of random variable and its relationship to the rate function
2. relationship between tails of random variable (support of mgf) and the steepness/domain of rate function

1.1 Cramér's and Sanov's theorem

Theorem 1.2 (Cramér). *Let (X_i) be i.i.d. $\mathbb{R}$-valued random variables with moment generating function $\varphi(t)$ defined on everywhere. Let $S_n = \sum_{i=1}^n X_i$. Then, for all $a > \mathbb{E}\, X_1$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > na) = -I(a) \tag{1.1}$$

with the rate function I be of the form

$$I(z) = \sup_{t\in\mathbb{R}} \{zt - \log\varphi(t)\}. \tag{1.2}$$

Proof Sketch. First, without loss of generality, we will consider the case where $a = 0$. This means that we want to show

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > 0) = -I(0) = \log \inf_{t\in\mathbb{R}} \varphi(t) := \log \rho. \tag{1.3}$$

In the case when $X_1 \le 0$, $\varphi(t)$ is decreasing and are straightfoward. We focus on the case where X_1 can take positive and negative values, giving a φ that is coercive.

Suppose the infimum of φ is achieved at τ, i.e., $\varphi(\tau) = \rho$. We will prove the upper and lower bounds separately. The upper bound follows from a clever application of Chebyshev's inequality:

$$\mathbb{P}(S_n > 0) = \mathbb{P}\left(e^{\tau S_n} > 1\right) \le \varphi(\tau)^n = \rho^n. \tag{1.4}$$

For the lower bound, let μ be the distribution function of X_1, and we introduce the *tilted measure* $\hat{\mu}$ with

$$\hat{\mu}(dx) = \frac{1}{\rho} e^{\tau x} \mu(dx). \tag{1.5}$$

Note that $\mu \ll \hat{\mu}$ as well by having a positive Radon-Nikodym derivative. Then, we can rewrite the desired probability in terms of the tilted measure

$$\mathbb{P}(S_n > 0) = \int_{\sum_i x_i > 0} \otimes_{i=1}^n \mu(dx_i) = \rho^n \int_{\sum_i x_i > 0} e^{-\tau \sum_i x_i} \otimes_{i=1}^n \hat{\mu}(dx_i). \tag{1.6}$$

It can be shown, via the Central Limit theorem for $\hat{S}_n$—S_n under the tilted measure—has no contribution to the exponential decay, i.e.,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\, e^{\tau \hat{S}_n} 1_{\hat{S}_n > 0} \ge 0. \tag{1.7}$$

□

Remark 1.3. Cramér's theorem would also hold even if φ only exists in a neighborhood of zero, however, it will require a different proof. To keep the proof as it, we can relax the condition to $\log\varphi$ being *steep*, that is, let D_φ be the domain for which φ is finite, then

$$\lim_{t\to\partial D_\varphi} |(\log\varphi)'(t)| \to \infty. \tag{1.8}$$

This is because $\log\varphi$ being steep, convex, and smooth guarantees that the Legendre transform is defined everywhere.

Definition 1.4. Let μ, ν be probability measures with $\nu \ll \mu$. Then, the relative entropy is defined as

$$\mathcal{R}(\nu\|\mu) = \mathbb{E}_\nu \log \frac{d\nu}{d\mu}. \tag{1.9}$$

Theorem 1.5 (Sanov)**.** *Let (X_i) be i.i.d. random variables taking value in the finite set $\Gamma = \{1, \ldots, r\}$ according to law ρ. We equip the space of probability measures $\mathcal{P}(\Gamma)$ with the total variation distance, and let the empirical measure $L_n = \sum_{i=1}^n \delta_{X_i}$. Then, for all $a > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a(\rho)^c) = - \inf_{\nu \in B_a(\rho)^c} \mathcal{R}(\nu\|\rho) \tag{1.10}$$

where $B_a(\rho)$ is the ball of radius a centered at ρ.

Proof Sketch. We do combinatorics! Let $k \in \Gamma^n$ be a sequence of possible outcome and $\nu_n(k)$ be the corresponding empirical measure. Then, for any n, there must be an empirical measure outside of $B_a(\rho)$ that is the most likely to occur; we call the probability of getting this empirical measure $Q_n(a)$, i.e.,

$$Q_n(a) = \max_{k:\nu_n(k)\in B_a(\rho)^c} n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!}. \tag{1.11}$$

Then, knowing that the total number of possible empirical measures is $\binom{n+r}{r-1}$, we first get the chain of inequality:

$$Q_n(a) \le \mathbb{P}(L_n \in B_a(\rho)^c) \le \binom{n+r}{r-1} Q_n(a). \tag{1.12}$$

Moreover, by Stirling's formula, $n^{-1} \log \nu_n(k) \longrightarrow -\mathcal{R}(\nu_n(k)\|\rho)$ for any empirical measure $\nu_n(k)$, including the one that achieves $Q_n(a)$. Combined with the facts that the set of empirical measures is dense in $\mathcal{P}(\Gamma)$ and continuity of $\nu \mapsto \mathcal{R}(\nu\|\rho)$, the statement follows. □

Remark 1.6. One way of linking Sanov's theorem and Cramér's theorem, at least in the finite-alphabet case, is via the convex dual. It can be shown that both the rate function in Cramér's theorem and the relative entropy (in the first argument) are convex. Since Cramér's theorem is about deviation from the mean, we fix $\mathbb{E} X_1 = z$. Then, convex duality says

$$\inf_{\nu\in\mathcal{P}(\Gamma)} \left\{ \sum_{s=1}^r \nu_s \log \frac{\nu_s}{\rho_s} : \sum_{s=1}^r s\rho_s = z \right\} = \sup_{t\in\mathbb{R}} \inf_{\nu\in\mathcal{P}(\Gamma)} \left\{ \sum_{s=1}^r \nu_s \log \frac{\nu_s}{\rho_s} + t\left(\sum_{s=1}^r s\nu_s - z \right) : \sum_{s=1}^r \nu_s = 1 \right\} \tag{1.13}$$

$$= \sup_{t\in\mathbb{R}} \left\{ tz - \log \sum_{s=1}^r e^{ts} \rho_s \right\}. \tag{1.14}$$

Formally, deriving the relationship between Cramér's theorem and Sanov's thoerem is done via the *contraction principle*, as demonstrated below.

Proposition 1.7 (Contraction Principle: Sanov to Cramér)**.** *Let (X_i) be i.i.d. random variables with the same set up as the previous (Sanov) theorem. For $\nu \in \mathcal{P}(\Gamma)$, denote $m_\nu = \sum_s s\nu_s$. Then, for all $a > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n \in B_a(m_\rho)^c \right) = - \inf_{z\in B_a(m_\rho)^c} I(z) \tag{1.15}$$

where the rate function takes the variational form

$$I(z) = \inf_{\nu\in\mathcal{P}(\Gamma)} \left\{ \mathcal{R}(\nu\|\rho) : m_\rho = z \right\}. \tag{1.16}$$

Proof. Let's first rewrite

$$\left\{ \frac{1}{n} S_n \in B_a(m_\rho)^c \right\} = \{ L_n \in \hat{B}_a(\rho)^c \} \tag{1.17}$$

where

$$\hat{B}_a(\rho) = \{\nu \in \mathcal{P}(\Gamma) : |m_\nu - m_\rho| \le a\}. \tag{1.18}$$

Then, adapting the proof to general open sets, it follows from Sanov that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n \in B_a(m_\rho)^c\right) = - \inf_{\nu \in \hat{B}_a(\rho)^c} \mathcal{R}(\nu \| \rho) = - \inf_{z \in B_a(m_\rho)^c} \inf_{\nu \in \mathcal{P}(\Gamma): m_\nu = z} \mathcal{R}(\nu \| \rho) \tag{1.19}$$

where the last equality follows from taking slices of $\hat{B}_a(\rho)^c$. □

Proposition 1.8 (Contraction Principle: finite to countable alphabets)**.** *Let (X_i) be i.i.d. random variables that has law ρ on $\mathbb{N}$. We equip the space the probability measures $\mathcal{P}(\mathbb{N})$ with the total variation distance. Then, for any $a > 0$, then letting $J(a) = \inf_{\nu \in B_a(\rho)^c} \mathcal{R}(\nu \| \rho)$, we have*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a(\rho)^c) \ge -J(a) \tag{1.20}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a(\rho)^c) \le -J(a^-). \tag{1.21}$$

Proof Sketch. We will do this via truncation/approximation. Let $\pi_N : s \mapsto s \wedge N$ and write $\pi_N \nu = \nu \circ \pi_N$. We again split into upper and lower bounds. Observing that we have uniform control over the total variation after truncation

$$0 \le d(\nu, \rho) - d(\pi_N \nu, \pi_N \rho) \le \frac{1}{2} \sum_{s=N}^{\infty} \rho_s + \nu_s \wedge \rho_s \le \sum_{s=N}^{\infty} \rho_s, \tag{1.22}$$

we can write the inequality

$$\mathbb{P}(d(\pi_N L_n, \pi_N \rho) > a) \le \mathbb{P}(d(L_n, \rho) > a) \le \mathbb{P}(d(\pi_N L_n, \pi_N \rho) > a - \delta) \tag{1.23}$$

for some $\delta > 0$. Apply Sanov's theorem for the finite alphabet case, we get

$$\begin{aligned} - \inf_{\nu \in \Gamma(\mathbb{N}) : d(\pi_N \nu, \pi_N \rho) > a} \mathcal{R}(\pi_N \nu \| \pi_N \rho) &\le \liminf \frac{1}{n} \log \mathbb{P}(d(L_n, \rho) > a) \\ &\le \limsup \frac{1}{n} \log \mathbb{P}(d(L_n, \rho) > a) \le - \inf_{\nu \in \Gamma(\mathbb{N}) : d(\pi_N \nu, \pi_N \rho) > a - \delta} \mathcal{R}(\pi_N \nu \| \pi_N \rho). \end{aligned} \tag{1.24}$$

□

1.2 Empirical measures of Markov Chains

1.3 Application of Sanov's theorem: interacting diffusions

We consider N particles diffusing down an energy landscape. Each particle X^i will be associate with an (random) medium/environment ω_i. The landscape is defined by the Hamiltonian

$$H_N(x, \omega) = \frac{1}{2N} \sum_{i,j=1}^{N} f(x_j - x_i; \omega_i, \omega_j) + \sum_{i=1}^{N} g(x_i; \omega_i) \tag{1.25}$$

where f is the pairwise potential and g is the internal field. Moreover, we will assume boundedness on f and g as well as their first and second derivatives. More importantly, we will need f to act symmetrically in the first argument, i.e., $f(x_j - x_i; \omega_i, \omega_j) = f(x_i - x_j; \omega_i, \omega_j)$ for all $1 \le i, j \le N$. Then, the particle "diffuse down the landscape" in the sense of being the solution to the SDE

$$\begin{cases} dX_t^i = -\frac{\partial H_N}{\partial x_i}(X_t; \omega) dt + dW_t^i, \\ X_0 \sim \lambda^{\otimes N}, \quad \omega \sim \mu^{\otimes N} \end{cases} \tag{1.26}$$

where $W = (W^i)_{i \ge 0}$ are independent Brownian motions.

Example 1.9 (Kuromoto)**.** The Kuromoto model is a common model of a large system of coupled oscillators. The particle will lie on the unit circle, i.e., $X_t^i \in [0, 2\pi]$ with periodic boundaries. For some (deterministic) coupling strength K, we can write the Hamiltonian as

$$H_N(x; \omega) = -\frac{1}{N} \sum_{i,j=1}^{N} K \cos(x_j - x_i) - \sum_{i=1}^{N} \omega_i x. \tag{1.27}$$

Without the coupling term, we see that each particle rotate at their own (random) inherit frequency ω. At the same time, the coupling term coerces the particles to synchronize with each other. We can show, using the tools to be developed, that in the case that μ is symmetric, there is a phase transition as K reaches a critical point K_c—if $K < K_c$, all particles are *incoherent* (position independently and uniformly distributed), and if $K > K_c$, a positive fraction of oscillators synchronize to the same frequency.

We are interested in establishing LDP in the large system-size limit for the empirical measure of the path and media jointly, i.e.,

$$L_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{(X_{0:T}^i, \omega_i)} \in \mathcal{P}(C[0:T] \times \mathbb{R}) \tag{1.28}$$

for some $T < \infty$.

Remark 1.10. We can consider different ensembles. The *quenched model* refers to fixing the media parameter, and observe that law, $\mathbb{P}_N^\omega$, of $(X_{0:T}^i)_{i\geq 0}$. On the other hand, when we take the media into account, that is, the law of $((X_{0:T}^i, \omega_i))_{i\geq 0}$, we call it the *annealed model.*

Remark 1.11. The quenched model has a well-known invariant distribution with density

$$x \mapsto \frac{1}{Z} e^{-H_N(x;\omega)} \tag{1.29}$$

where Z is the normalizing constant; this is the *Gibbs distribution.* Under the inner-product space induced by this invariant distribution, the process is reversible.

We will present the LDP two ways: the first is an easy consequence of the theorems we've established before, namely, Sanov's theorem and the tilted LDP. However, the rate function is not interpretable in this form, so we resort to the second derivation and embrace the idea of a typical particle. We start with the not-so-handsome-looking one.

Lemma 1.12. *Let $\mathbb{P}_N^\omega$ be the annealed law and $\mathbb{W}_N$ be the law of the Brownian motion driving the system, i.e., the system if $H \equiv 0$. Then, there exists a functional $F : \mathcal{P}(C[0 : T] \times \mathbb{R}) \to \mathbb{R}$ such that the Radon-Nikodym derivative can be written as*

$$\frac{d\mathbb{P}_N^\omega}{d\mathbb{W}_N}(X_{0:T}) = e^{NF(L_N)} \tag{1.30}$$

where the empirical measure is on the trajectories of each particle of the sample path $X_{0:T}$ with corresponding media ω.

Proof. Via Girsanov, the Radon-Nikodym derivative can be written down as

$$\frac{d\mathbb{P}_N^\omega}{d\mathbb{W}_N}(X_{0:T}) = \exp\left(-\sum_{i=1}^{N} \int_0^T \frac{\partial H_N}{\partial x_i}(X_t; \omega) dX_t^i - \frac{1}{2} \sum_{i=1}^{N} \int_0^T \left(\frac{\partial H_N}{\partial x_i}(X_t; \omega) \right)^2 dt \right). \tag{1.31}$$

Recall that, under $\mathbb{W}_N$, X^i's are independent Brownian motions. Therefore, by Ito, we can rewrite the second term as

$$\sum_{i=1}^{N} \int_0^T \frac{\partial H_N}{\partial x_i}(X_t; \omega) dX_t^i = H(X_T) - H(X_0) - \frac{1}{2} \sum_{i=1}^{N} \int_0^T \frac{\partial^2 H_N}{\partial x_i^2}(X_t; \omega) dt. \tag{1.32}$$

Using the symmetry of f, that is,

$$\frac{\partial H_N}{\partial x_i}(X_t;\omega) = \frac{\partial}{\partial x_i}\left(\frac{1}{2N}\sum_{j=1}^{N} f(x_i - x_j) + f(x_j - x_i)\right) = \frac{1}{N}\sum_{j=1}^{N} f'(x_i - x_j), \tag{1.33}$$

the exponent of the Radon-Nikodym derivative looks like

$$\begin{aligned}
&-(H(X_T) - H(X_0)) + \frac{1}{2}\sum_{i=1}^{N}\int_0^T \frac{\partial^2 H_N}{\partial x_i^2}(X_t;\omega)dt - \frac{1}{2}\sum_{i=1}^{N}\int_0^T \left(\frac{\partial H_N}{\partial x_i}(X_t;\omega)\right)^2 dt \\
&= -\left(\frac{1}{2N}\sum_{i,j=1}^{N} f(X_T^i - X_T^j;\omega_i,\omega_j) - f(X_0^i - X_0^j;\omega_i,\omega_j) + \sum_{i=1}^{N} g(X_T^i;\omega_i) - g(X_0^i;\omega_i)\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{N}\int_0^T \frac{1}{N}\sum_{j=1}^{N} f''(X_t^i - X_t^j;\omega_i,\omega_j) - g''(X_t^i;\omega_i)dt \\
&\quad - \frac{1}{2}\sum_{i=1}^{N}\int_0^T \left(\frac{1}{N}\sum_{j=1}^{N} f'(X_t^i - X_t^j;\omega_i,\omega_j) - g'(X_t^i;\omega_i)\right)^2 dt.
\end{aligned} \tag{1.34}$$

Now, thinking of each empirical average as integrating over empirical measures and adding a factor of N in front of each term yields $d\mathbb{P}_N^{\omega}/d\mathbb{W}_N = e^{NF(L_N)}$. □

Theorem 1.13. *The sequence of empirical measures $(L_N)_N \subset \mathcal{P}(C[0:T]\times\mathbb{R})$ satisfies an LDP with rate N and rate function*

$$I(Q) = \mathcal{R}(Q\|W\times\mu) - F(Q) \tag{1.35}$$

where W is the Wiener measure.

Proof. First, we know that, under $\mu^{\otimes N}(d\omega)\mathbb{W}_N(dx_{0:T})$, we can apply Sanov's theorem and deduce that (L_N) satisfies an LDP with rate N and rate function $\mathcal{R}(Q\|W\times\mu)$. Now, we can tilt $\mathbb{W}_N$ by $e^{NF(L_N)}$ to get the quenched law $\mathbb{P}_N^{\omega}$. Note that F is continuous and bounded (using the boundedness of f and g and their derivatives) in the topology of weak convergence of measures. Thus, using the tilted LDP, we get that, under the annealed model $\mu^{\otimes N}(d\omega)\mathbb{P}_N^{\omega}(dx_{0:T})$, (L_N) satisfies an LDP with rate N and rate function $\mathcal{R}(Q\|W\times\mu) - F(Q)$. □

Looking at the proof, we see that we get the LDP practically for free—particularly from boundedness assumptions as well as an interaction kernel that scales like $1/N$—using two of the previous theorems. The downside is also clear: it is difficult to interpret the rate function, especially the functional F. As an attempt to simplify this, let's take a philosophical detour. From the large deviation principle, we know that there is a deterministic measure on $C[0:T]\times\mathbb{R}$—the measure for which the rate function is zero—that the empirical measure converge to. Such measure will describe the average media parameter, as well as the conditional law of an average/typical particle given the media. We will proceed to indulge in the idea of the typical particle and derive properties of the system in this *mean-field* limit.

Suppose $Q \in \mathcal{P}(C[0:T]\times\mathbb{R})$ is the law of the typical particle and, additionally, we pick our favorite out of the litter. Suppose that our particle is wandering in media ω. Then, our particle should follow the SDE

$$dX_t = \beta^{\omega,\pi_t Q}(X_t)dt + dW_t, \quad X_0 \sim \lambda \tag{1.36}$$

where π_t is the projection onto the time-t marginal, i.e., $\pi_t Q \in \mathcal{P}(\mathbb{R}\times\mathbb{R})$ and $\beta^{\omega,q}$ diffuses down the energy landscape with respect to the mean-field effect

$$\beta^{\omega,q}(x) = -\int f'(y - x;\omega,\pi) - g'(x)q(dy,d\pi). \tag{1.37}$$

We will denote that quenched law of the unique strong solution to this SDE by $\mathbb{P}^{\omega,Q}$. Of course, as much as we'd like our child to be the best, our favorite particle of the litter is nothing but another typical particle, so the law of X has to be the quenched law derived from Q. However, this is not a priori obvious and the existence of such process will be established through the LDP below.

Theorem 1.14. *For any $Q \in \mathcal{P}(C[0:T] \times \mathbb{R})$ and let $\mathbb{P}^Q(dx_{0:T}, d\omega) = \mathbb{P}^{\omega,Q}(dx_{0:T})\mu(d\omega)$. Then, $(L_N)_N$ satisfies an LDP with rate n and rate function*

$$I(Q) = \mathcal{R}(Q \| \mathbb{P}^Q). \tag{1.38}$$

Proof. This theorem relies on the previous less interpretable LDP. Notice that by the chain rule of relative entropy,

$$\mathcal{R}(Q\|\mathbb{P}^Q) = \int \log \frac{dQ}{d\mathbb{P}^{\omega,Q}\mu} dQ = \int \log \frac{dQ}{dW \times \mu} dQ - \int \log \frac{d\mathbb{P}^{\omega,Q}}{dW} dQ. \tag{1.39}$$

So, we have to verify that $F(Q) = \int \log \frac{d\mathbb{P}^{\omega,Q}}{dW} dQ$. First, by Girsanov, we find the (log of) the Radon-Nikodym derivative to be

$$\log \frac{d\mathbb{P}^{\omega,Q}}{dW}(x_{0:T}) = \int_0^T \beta^{\omega,\pi_t Q}(x_t)dx_t - \frac{1}{2}\int_0^T \left(\beta^{\omega,\pi_t Q}(x_t)\right)^2 dt. \tag{1.40}$$

We start with the first term. Recall that f is symmetric and f' is odd, so

$$\begin{aligned}
&\int\int\int_0^T f'(y_t - x_t; \omega, \pi)dx_t Q(dx_{0:T}, d\omega)Q(dy_{0:T}, d\pi) \\
&= \frac{1}{2}\int\int\int_0^T f'(y_t - x_t; \omega, \pi)d(x_t - y_t)Q(dx_{0:T}, d\omega)Q(dy_{0:T}, d\pi) && (1.41)\\
&= \frac{1}{2}\int\int\left(\int_0^T f''(y_t - x_y; \omega, \pi)dt - f(y_0 - x_0; \omega, \pi) + f(y_T - x_T)\right)Q(dx_{0:T}, d\omega)Q(dy_{0:T}, d\pi) && (1.42)
\end{aligned}$$

where the last equality is by Ito's formula for the semi-martingale $x_t - y_t$, i.e.,

$$f(y_T - x_T) - f(y_0 - x_0) = \int_0^T f'(y_t - x_t)d(x_t - y_t) + \int_0^T f''(y_t - x_t)dt. \tag{1.43}$$

Since the integrands are bounded, using Fubini and repeating similar steps as above for g, we get

$$\begin{aligned}
&\int\int_0^T \beta^{\omega,\pi_t Q}(x_t)dx_t Q(dx_{0:T}, d\omega) \\
&= -\int\int_0^T \left(\int f'(y_t - x_t; \omega, \pi)Q(dy_{0:T}, d\pi) + g'(x_t; \omega)\right)dx_t Q(dx_{0:T}, d\omega) && (1.44)\\
&= \frac{-1}{2}\int\int\left(\int_0^T f''(y_t - x_y; \omega, \pi)dt - f(y_0 - x_0; \omega, \pi) + f(y_T - x_T)\right)Q(dx_{0:T}, d\omega)Q(dy_{0:T}, d\pi) \\
&\qquad - \int\int\left(-\frac{1}{2}\int_0^T g''(x_t; \omega)dt + g(x_T; \omega) - g(x_0; \omega)\right)Q(dx_{0:T}, d\omega) && (1.45)
\end{aligned}$$

The second term is a bit less work. Simply expanding gives

$$\begin{aligned}
&-\frac{1}{2}\int\int_0^T \left(\beta^{\omega,\pi_t Q}(x_t)\right)^2 dt Q(x_{0:T}, d\omega) \\
&= -\frac{1}{2}\int_0^T \left(\int f'(x_t - y_t; \omega, \pi)\pi_t Q(dy_t, d\pi) - g'(x_t; \omega)\right)^2 dt && (1.46)\\
&= -\frac{1}{2}\int\int_0^T \left(\int f'(x_t - y_t; \omega, \pi)Q(dy_{0:t}, d\pi) - g'(x_t; \omega)\right)^2 dt Q(x_{0:T}, d\omega). && (1.47)
\end{aligned}$$

Pattern match the above to equation (1.34) gives the result. □

It turns out, after gruesome calculations, that our "typical particle" way of thinking is fruitful! As rate functions always have a zero, we know that there must be a measure $Q = P^Q$ that describes a typical particle, and the evolution by the average drift $\beta^{\omega,Q}$ must yield a law that agrees with Q (averaged over the media variable ω). We call this the *McKean-Vlasov diffusion* and we remark a few of its properties below.

Theorem 1.15 (McKean-Vlasov). *Let $Q \in \mathcal{P}(C[0:T] \times \mathbb{R})$ be the fixed point $Q = P^Q$, disintegrate $Q(dx_{0:T}, d\omega) = \nu(d\omega)Q^\omega(dx_{0:T})$. Assume that $X_0 \sim \lambda$ has a density with respect to Lebesgue measure and has finite p-th moment for some $p > 1$. Then, the following holds:*

1. *$\nu = \mu$,*
2. *Q^ω is the law of the time-inhomogenious Markov process*

$$dX_t = \beta^{\omega,\pi_t Q}(X_t)dt + dW_t,$$

3. *let $q_t^\omega = \pi_t Q^\omega$ be the time-marginal of the process, then it is the weak solution of the McKean-Vlasov equation*

$$\begin{cases} \partial_t q_t^\omega = \mathcal{L}^\omega q_t^\omega, \\ q_0^\omega = \lambda \end{cases} \tag{1.48}$$

where

$$\mathcal{L}^\omega q_t^\omega = -\frac{\partial}{\partial x}(\beta^{\omega,q_t} q_t^\omega) + \frac{1}{2}\frac{\partial^2}{\partial x^2} q_t^\omega, \tag{1.49}$$

4. *the diffusion process has the generator*

$$L_t^\omega = \beta^{\omega,q_t}\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2}. \tag{1.50}$$

Notice that, since $\beta^{\omega,q}$ depends on the law of the process, this diffusion is heavily nonlinear. But on the upside, we've reduced the dimensionality of the system from N (in the limit as $N \to \infty$) to 1. However, finding Q is usually difficult.

We will conclude the section filling in the analysis of the Kuromoto model. Let q^ω be the density of the typical particle, which must satisfy the boundary condition $q_t^\omega(0) = q_t^\omega(2\pi)$. We will define the *order parameter*

$$r_t e^{i\psi_t} = \int_{\mathbb{R}} \int_0^{2\pi} e^{ix} q_t^\omega(x) dx \mu(d\omega) \tag{1.51}$$

where we call r_t the *phase coherence* and ψ_t the *average phase*. We can express the drift of the McKean-Vlasov diffusion in terms of the order paramters:

$$\beta_t^{\omega,q_t} = \int_{\mathbb{R}} \int_0^{2\pi} (K\sin(y-x) + \omega)\, q^\pi(dy)\mu(d\pi) = Kr_t\sin(\psi_t - x) + \omega. \tag{1.52}$$

For simplicity, let's assume μ is symmetric so that the imaginary part vanishes and $\psi_t = 0$. Moreover, we will try to find a stationary solution. From the McKean-Vlasov equation and taking the time evolution to be zero, we will find that $q_t^\omega := q$ has to satisfy the second-order ODE

$$q'' - (Kr\sin x + \omega)q' + (Kr\cos x)q = 0. \tag{1.53}$$

First, when $r = 0$, we can easily solve the ODE

$$q'' - \omega q' = 0$$

with the boundary condition $q(0) = q(2\pi)$ yields the solution that q is a constant. This corresponds to the incoherent phase where all oscillators are uniformly distributed and independent of each other.

Life becomes much harder when $r > 0$. First, we would have to solve the second-order ODE (up to intergrating factors), so

$$q^\omega(x) = \frac{1}{Z^{\omega,r}} A^{\omega,r}(x) \tag{1.54}$$

where

$$A^{\omega,r}(x) = B^{\omega,r}(x)\left(e^{4\pi\omega}\int_0^{2\pi}\frac{dy}{\mathbb{B}^{\omega,r}} + (1 - e^{4\pi\omega}\int_0^{x}\frac{dy}{B^{\omega,r}(y)}\right), \tag{1.55}$$

$$B^{\omega,r}(x) = \exp(2Kr\cos x + 2\omega x). \tag{1.56}$$

If we look back at the order parameters, we can plug our r-dependent solutions back in and obtain a *consistency relation*

$$r = \Phi_\mu(r) = \int_{\mathbb{R}}\frac{1}{Z^{\omega,r}}\left(\int_0^{2\pi}A^{\omega,r}(x)\cos x dx\right)\mu(d\omega). \tag{1.57}$$

The existence of a solution relies on whether the consistency relation is satisfied. If we study the map $r \mapsto \Phi(r)$, we can observe that

1. the map is continuous with $\Phi_\mu(0) = 0$ and $\Phi_\mu(r) \to 1$ as $r \to \infty$,
2. if μ is unimodal, $\Phi'_\mu(0) = K/K_c$, $\Phi''_\mu(0) = 0$, and $\Phi'''_\mu(0) < 0$ where

$$K_c^{-1} = \int_{\mathbb{R}}\frac{1}{1+4\omega^2}\mu(d\omega). \tag{1.58}$$

Therefore, for $K/K_c > 1$, or $K > K_c$, we know that $\Phi_\mu(r) > r$ for small enough r and $\Phi_\mu(r) < r$ for $r > 1$. Therefore, by continuity, there must exist a fixed point $r = \Phi_\mu(r)$ and a synchronized solution exists (though we know nothing about uniqueness). When $K/K_c < 1$, it is unclear as to whether a solution exist as we don't have exact control over the concavity.

2 General Theory of Large Deviation

2.1 Rate functions and the Laplace principle

2.2 Gärtner-Ellis theorem So far, concrete examples of the large deviation principle has been based on strong model assumptions, e.g., i.i.d. samples or Markov chain. Here, we hope to generalize the previous examples which will be done via careful convex analysis. Let $Z_n : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be a sequence of random variables and let φ_n denote its moment generating functions.

Assumption 2.1. *Below is the running assumption throughout the subsection:*

1. $\lim_{n\to\infty}\frac{1}{n}\log\varphi_n(nt) = \Lambda(t)$ *exists, and*
2. $0 \in \operatorname{int} D_\Lambda$.

We're interested in establishing large deviation for the family of measures $P_n(\cdot) = \mathbb{P}(Z_n \in \cdot)$ through the convex dual $\Lambda^*(x) = \sup_{t\in\mathbb{R}}\{\langle t, x\rangle - \Lambda(t)\}$. Of course, we must first make sure that Λ^* makes sense as it is not a priori obvious with Λ being a limit.

Lemma 2.2. *Under Assumption 2.1, Λ is convex and bounded below and Λ^* is a good rate function and convex.*

Proof. Since 0 remains in the domain of Λ and $\log\varphi_n$ are convex functions, Λ is convex and bounded from below. Immediately, this implies that the convex dual Λ^* is also convex and lower semi-continuous. Lastly, to show that Λ^* has compact level sets, it suffices to show that the level sets are bounded. Since $0 \in \operatorname{int} D_\Lambda$, there is a $\delta > 0$ such that $B_\delta(0) \subset \operatorname{int} D_\Lambda$. So,

$$\Lambda^*(x) \geq \sup_{t\in B_\delta(0)}\langle t, x\rangle - \Lambda(t) \geq \delta|x| - \sup_{t\in B_\delta(0)}\Lambda(t). \tag{2.1}$$

Since the latter function has bounded level sets, Λ^* must also have bounded level sets. Combined with lower semi-continuity of Λ^*, we get that Λ^* is a good rate function. □

Remark 2.3. The assumption that Λ—as the pointwise limit of functions—exists can be thought of a requirement on the "strength" of dependence between Z_n's. First, if $(X_n)_{n\geq 0}$ be i.i.d. and $Z_n = \frac{1}{n}\sum_{k=1}^n X_k$, then

$$\frac{1}{n}\log\varphi_n(nt) = \frac{1}{n}\log\left(\mathbb{E}\, e^{tX_1}\right)^n = \mathbb{E}\, e^{tX_1} \tag{2.2}$$

as we have with Cramér. In fact, we can let the samples be a bit more dependent: consider $(X_n)_{n\geq 0}$ as a (strong-sense) stationary, mean-zero Gaussian process whose correlation decays fast enough

$$\sum_{k=1}^{\infty} |C_k| := \sum_{k=1}^{\infty} |\mathbb{E}\, X_1 X_k| < \infty. \tag{2.3}$$

Again, let Z_n be the sample mean. Then, we can calculate

$$\frac{1}{n}\log\varphi_n(nt) = \frac{1}{n}\log e^{\frac{t^2}{2}\sum_{k=1}^n C_k(n-|k|)} = \frac{t^2}{2}\sum_{k=1}^n C_k\left(1-\frac{|k|}{n}\right) \tag{2.4}$$

which has a well-defined limit. However, this will collapse if we go too extreme: let Y be non-degenerate and $X_n \equiv Y$ for all n. Then,

$$\frac{1}{n}\log\varphi_n(nt) = \frac{1}{n}\log\mathbb{E}\, e^{ntY} \longrightarrow \infty \tag{2.5}$$

as $n \longrightarrow \infty$ since the log-moment generating function grows super-linearly.

As hinted before, the theorem relies on a fair chunck of convex analysis. To obtain a proper lower bound later, we need to restrict our attention to areas where the probability escapes super-linearly (in log scale). We formalize this below.

Definition 2.4. A point $x \in \mathbb{R}^d$ is exposed for Λ^* if there is a $t \in \mathbb{R}^d$ such that

$$\Lambda^*(y) - \Lambda^*(x) > \langle y - x, t\rangle \tag{2.6}$$

for any $y \neq x$. We say that t is normal to an exposing hyperplane for x.

Now, we're ready to present the theorem.

Theorem 2.5 (Gärtner-Ellis). *Given Assumption 2.1 and let E be the set of exposed points of Λ^* belonging to* $\operatorname{int} D_\Lambda$*, the sequence of (P_n) satisfies a LDP with rate n and rate function Λ^* on the exposed points, i.e.,*

1. $\limsup_{n\to\infty} \frac{1}{n}\log P_n(C) \leq -\inf_{x\in C}\Lambda^*(x)$ *for closed $C \subset \mathbb{R}^d$, and*
2. $\liminf_{n\to\infty} \frac{1}{n}\log P_n(O) \leq -\inf_{x\in O\cap E}\Lambda^*(x)$ *for open $O \subset \mathbb{R}^d$.*

Proof. The proof will proceed like that of Cramér's theorem—upper bound is established via an exponential Chebyshev inequality, and the lower bound via exponential tilting—though requiring more intricate analysis throughout.

Upper bound for compact sets We first prove the statement for compact sets. Pick your favorite $\delta > 0$ and shrink and cleverly truncate Λ^* be defining

$$\Lambda^*_\delta = \min\left\{\Lambda^* - \delta, \frac{1}{\delta}\right\}. \tag{2.7}$$

Now, for every $x \in \mathbb{R}^d$, we can pick a δ-optimizer $t_x \in \mathbb{R}^d$ such that

$$\langle x, t_x\rangle - \Lambda(t_x) \geq \Lambda^*(x) - \delta \geq \Lambda^*_\delta(x) \tag{2.8}$$

while finding a neighborhood A_x such that

$$\inf_{y\in A_x}\langle y - x, t_x\rangle \geq -\delta. \tag{2.9}$$

By Chebyshev while exponentiating by a factor of n, we get that

$$P_n(A_x) \leq \mathbb{P}(\langle Z_n - x, t_x\rangle \geq -\delta) \leq e^{\delta n}\varphi_n(nt_x)e^{-n\langle x,t_x\rangle}. \tag{2.10}$$

Now, take a compact set K and cover it by $\bigcup_{x\in K} A_x \supset K$. By compactness, we only need to finitely many A_x's $\bigcup_{i=1}^{N} A_{x_i}$. So, we can estimate

$$\frac{1}{n}\log P_n(K) \leq \frac{1}{n}\log\left(N \max_{i=1,\dots,N} P_n(A_{x_i})\right) \leq \frac{1}{n}\log N + \delta - \min_{i=1,\dots,N}\left\{\langle x_i, t_i\rangle - \frac{1}{n}\log\varphi_n(nt_{x_i})\right\}. \tag{2.11}$$

Take $n \longrightarrow \infty$ gives

$$\limsup_{n\to\infty}\frac{1}{n}\log P_n(K) \leq \delta - \min_{i=1,\dots,N}\left\{\langle x_i, t_i\rangle - \Lambda(t_{x_i})\right\} \leq \delta - \min_{i=1,\dots,N}\Lambda^*_\delta(t_{x_i}) \leq \delta - \inf_{x\in K}\Lambda^*_\delta(x). \tag{2.12}$$

Taking $\delta \downarrow 0$ completes the upper bound for compact sets.

Upper bound for general closed sets To start by showing the sequence (P_n) is exponentially tight. Consider a sequence of growing cubes $[-N,N]^d \subset \mathbb{R}^d$ and let $(e_i)_{i=1}^d$ denote the canonical basis of $\mathbb{R}^d$. Since $0 \in \operatorname{int} D_\Lambda$, there is $\delta > 0$ such that $\pm\delta e_i \in \operatorname{int} D_\Lambda$ for all i. From Chebyshev while exponentiating by $n\delta$, we get the following pair of inequality:

$$\mathbb{P}(Z_{n,i} \geq N) \leq e^{-n\delta N}\varphi_n(n\delta e_i), \;\; \mathbb{P}(Z_{n,i} \leq -N) \leq e^{-n\delta N}\varphi_n(-n\delta e_i). \tag{2.13}$$

By the same type of estimate as before, we get

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}(Z_n \in \mathbb{R}^d \setminus [-N,N]^d) \leq -\delta N + \max_{i=1,\dots,d}\max\left\{\Lambda(\delta e_i), \Lambda(-\delta e_i)\right\}. \tag{2.14}$$

Taking $N \longrightarrow \infty$, the above drops to $-\infty$ and exponential tightness is established.

Now, take any closed $C \subset \mathbb{R}^d$, notice that $C \cap [-N,N]^d$ is compact and

$$\inf_{x\in C}\Lambda^*(x) = \lim_{N\to\infty}\inf_{x\in C\cap[-N,N]^d}\Lambda^*(x). \tag{2.15}$$

So, we decompose C into inside and outside the cube and use the results established for compact set to write

$$\limsup_{n\to\infty}\frac{1}{n}\log P_n(C) \leq \limsup_{n\to\infty}\frac{1}{n}\left(\log P_n(C\cap[-N,N]^d) + P_n(\mathbb{R}^d\setminus[-N,N]^d)\right) \tag{2.16}$$

$$\leq \max\left\{-\inf_{x\in C\cap[-N,N]^d}\Lambda^*(x), \limsup_{n\to\infty}\frac{1}{n}\log P_n(\mathbb{R}^d\setminus[-N,N]^d)\right\} \tag{2.17}$$

$$\longrightarrow -\inf_{x\in C}\Lambda^*(x) \tag{2.18}$$

as $N \longrightarrow \infty$ since the second term goes to $-\infty$ by exponential tightness.

Lower bound It is enough to show that for any ball $B_\epsilon(x)$

$$\lim_{\epsilon\downarrow 0}\liminf_{n\to\infty}\frac{1}{n}\log P_n(B_\epsilon(x)) \geq -\Lambda^*(x) \tag{2.19}$$

as we can always put little balls inside open sets and optimize over the placement of the ball later. We proceed by exponential tilting.

Fix any $x \in E$ and let $\tau \in \operatorname{int} D_\Lambda$ be an exposing hyperplane for x. So, we introduce a new measure $\hat{P}_n$ with Radon-Nikodymn derivative

$$\frac{d\hat{P}_n}{dP_n}(y) = \frac{1}{\varphi_n(n\tau)}e^{n\langle y,\tau\rangle}. \tag{2.20}$$

Then, doing a change in measure gives:

$$\frac{1}{n}\log P_n(B_\epsilon(x)) = \frac{1}{n}\log\varphi_n(n\tau) + \frac{1}{n}\log\int_{B_\epsilon(x)} e^{-n\langle y,\tau\rangle}\hat{P}_n(dy) \tag{2.21}$$

$$\geq \frac{1}{n}\log\varphi_n(n\tau) - \langle x,\tau\rangle - \epsilon|\tau| + \frac{1}{n}\log\hat{P}_n(B_\epsilon(x)) \tag{2.22}$$

Notice that, taking the limit, the first three terms combined gives the lower bound desired. So, we need to show that the last term goes to zero, i.e., the event $B_\epsilon(x)$ is not rare in the tilted measure—or equivalently, $\mathbb{R}^d \setminus B_\epsilon(x)$ is rare. First, notice that

$$\lim_{n\to\infty}\frac{1}{n}\log\hat{\varphi}_n(nt) = \lim_{n\to\infty}\frac{1}{n}\log\frac{\varphi_n(n(t+\tau))}{\varphi_n(n\tau)} = \Lambda(t+\tau) - \Lambda(\tau), \tag{2.23}$$

which means that

$$\hat{\Lambda}^* = \sup_t\{\langle x,t\rangle - \Lambda(t+\tau) + \Lambda(\tau)\} = \sup_t\{\langle x,t+\tau\rangle - \Lambda(t+\tau)\} - \langle x,\tau\rangle + \Lambda(\tau) = \Lambda^*(x) - \langle x,\tau\rangle + \Lambda(\tau). \tag{2.24}$$

Since Λ^* is a good rate function and, from before, we can obtain a large deviation upper bound on the set $\mathbb{R}^d \setminus B_\epsilon(x)$ and suppose the infimum of $\hat{\Lambda}^*$ is achieved on x_0. However, since x is in E and τ is an exposing hyperplane for Λ^*, we have

$$\hat{\Lambda}^*(x_0) = \Lambda^*(x_0) - \langle x_0,\tau\rangle + \Lambda(\tau) \geq \Lambda^*(x_0) - \langle x_0,\tau\rangle - \Lambda^*(x) + \langle x,\tau\rangle > 0. \tag{2.25}$$

Therefore, for any fixed ϵ, the rate outside of the ball is less than 0, meaning that the probability outside decays to zero in the limit. However, since probability must sum up to one, we know that the probability inside must grow to one in the limit. Sending $\epsilon \downarrow 0$ completes the proof. □

Remark 2.6. The theorem can be strengthened to a full LDP if we add the constraint that 1) Λ is lower semi-continuous, 2) Λ is differentiable in $\operatorname{int} D_\Lambda$, and 3) either $D_\Lambda = \mathbb{R}^d$ or ∂D_Λ is steep. In this case, optimizing Λ^* over any open set is the same as optimizing over the relative interior of Λ^*, which is a subset of points with an exposing hyperplane.

Remark 2.7. To see why we need the exposing hyperplane (strong convexity) formalism, take $X_n = Y \sim \frac{1}{2}\delta_{\{-1\}} + \frac{1}{2}\delta_{\{1\}}$.

3 Large Deviation via Weak Convergence

3.1 Sanov's theorem revisited

3.2 Small-noise limit of diffusion processes

3.3 Small-noise limit of SDEs driven by jumps

- Finite time horizon $0 < T < \infty$, $(\Omega, \mathcal{F}, \mathbb{P})$ probability space, filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying usual conditions.
- Define $\mathcal{F}_t$-Poisson process $N : \Omega \to D[0,1]$ such that N_t is adapted for all $t \in [0,T]$ with independent increments, and $N_t - N_s \sim \text{Poisson}(1)$ for $s < t$.
- For $\theta > 0$, let N^θ denote a Poisson process with intensity θ.
- Controlled dynamics:
  - Consider a different probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}}, \{\bar{\mathcal{F}}_t\}_{t\geq 0})$ with filtration satisfying usual conditions.
  - Let $\mathcal{A} = \{\varphi : [0,t] \times \bar{\Omega} \to [0,\infty)\}$ be the set of predictable processes with $\int_0^t \varphi(s)ds < \infty$ a.s.

- The controlled process $N^{\theta\varphi}$ is a Poisson process with intensity $\theta\varphi$ such that for all bounded $f : [0,\infty) \to [0,\infty)$,

$$f(N^{\theta\varphi}(t)) - f(0) - \theta \int_0^t \varphi(s) \left(f(N^{\theta\varphi}(s) + 1) - f(N^{\theta\varphi}(s)) \right) ds \tag{3.1}$$

  is a $\bar{\mathcal{F}}_t$-martingale.
- Define the set of restricted controls

$$S_M = \left\{ \phi : [0,1] \to \mathbb{R}^+ : \int_0^1 \ell(\phi(s))ds \leq M \right\} \tag{3.2}$$

  where $\ell(x) = x \log x - x + 1$. Let $\mathcal{A}_{b,M}$ be the set of controls for which $\varphi \in \mathcal{A}$, $\varphi(\omega) \in S_M$ for all $\omega \in \bar{\Omega}$ and there is a $K < \infty$ such that $K^{-1} \leq \phi \leq K$.

These inequalities for ℓ will turn out to be useful for getting estimates.

Lemma 3.1. *For $a, b \geq 0$ and $c \geq 1$, we have*

$$ab \leq e^{ca} + \frac{\ell(b)}{c}, \quad b \leq e + \ell(b). \tag{3.3}$$

We're interested in establishing an LDP for the small-noise limit of SDEs driven by Poisson processes:

$$dX^n(t) = b(X^n(t))dt + \frac{1}{n}\sigma(X^n(t^-))dN^n(t), \quad X^n(0) = x. \tag{3.4}$$

For the existence of ODE limits as well as simplicity, we assume the following.

Assumption 3.2. *We assume that b and σ are Lipschitz and bounded, i.e., there exists $C > 0$ such that*

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq C|x - y|, \quad |b(x)| + |\sigma(x)| \leq C \tag{3.5}$$

for all $x, y \in \mathbb{R}$.

We want to prove the following.

Theorem 3.3. *The collection $(X^n)_{n\geq 0}$ satisfies an LDP with rate function*

$$I(\psi) = \inf_{\gamma \in U_\psi} \left\{ \int_0^1 \ell(\gamma(t))dt \right\} \tag{3.6}$$

where $U_\psi = \left\{ \gamma \in L^1([0,1]) : \psi(\cdot) = x + \int_0^\cdot b(\psi(s))ds + \int_0^\cdot \sigma(\psi(s))\gamma(s)ds \right\}$.

The proof structure will follow the previous examples: establishing variational lower bounds via compactness arguments and upper bounds via picking a near optimal control. Central to the proof, again, is the representation formula.

Proposition 3.4 (Representation formula for Poisson processes)**.** *Let $G : D[0,1] \to \mathbb{R}$ be a bounded and Borel-measurable and let $\theta \in (0,\infty)$. Then,*

$$-\log \mathbb{E}\, e^{-G(N^\theta)} = \inf_{\varphi \in \mathcal{A}} \mathbb{E}\, G(N^{\theta\varphi}) + \theta \int_0^1 \ell(\varphi(s))ds. \tag{3.7}$$

Moreover, for any $\delta > 0$, there is an $M = M(\|G\|_\infty, \delta)$ such that for all θ,

$$-\frac{1}{\theta} \log \mathbb{E}\, e^{-\theta G(N^\theta)} \geq \inf_{\varphi \in \mathcal{A}_{b,M}} \mathbb{E}\, G(N^{\theta\varphi}) + \int_0^1 \ell(\varphi(s))ds - \delta. \tag{3.8}$$

Proof of Theorem 3.3. We split the proof into four sections. First, we identify the variational form of interest for the LDP. Then, we prepare ourselves for compactness arguments by establishing tightness of the state process and controls. Finally, we prove the upper and lower bound for Laplace principle separately.

Applying the variational formula To prove the Laplace principle, we are interested in functionals for the state process X^n. However, similar to the Brownian motion case, there is always a unique strong solution to the SDE such that there is a measurable map $\mathcal{G}^n : D[0,1] \to D[0,1]$ such that $X^n = \mathcal{G}^n(N^n)$. Therefore, for the rest of the proof, we will fix a bounded continuous $F : D[0,1] \to \mathbb{R}$, and since $F \circ \mathcal{G}$ is still bounded and measurable, we can apply the variational formula:

$$\frac{-1}{n} \log \mathbb{E}\, e^{-nF(X^n)} = \inf_{\varphi \in \mathcal{A}} \left\{ \mathbb{E}\, F(X^n) + \int_0^1 \ell(\varphi(t))dt \right\} \tag{3.9}$$

$$\geq \inf_{\varphi \in \mathcal{A}_{b,M}} \mathbb{E}\, F(X^n) + \int_0^1 \ell(\varphi(s))ds - \delta. \tag{3.10}$$

The boundedness of controls is important for establishing tightness, which we are moving onto now.

Establishing tightness Let's first consider the controls. The trick to establishing tightness for the set of control realizations S_M is to think of it as a subset of the space of measures and equip it with the topology of weak convergence. That is, for each element $\gamma \in S_M$, we associate a measure ν^γ on $[0,1]$ equipped with the Borel σ-algebra such that $v^\gamma(dx) = \gamma(x)dx$, i.e.,

$$\gamma \in S_M \Leftrightarrow \int_0^1 \ell(\gamma)ds = \mathcal{R}(\nu^\gamma \| m) \leq M$$

where m is the Lebesgue measure. Then, compactness of S_M follows from compactness of the support and lower-semicontinuity of relative entropies; tightness of $\{\varphi^n\} \subset \mathcal{A}_{b,M}$ then follows from compactness of S_M.

Now, let $\bar{X}^n$ denote the controlled process, that is, it satisfies the SDE with initial condition

$$d\bar{X}^n(t) = b(\bar{X}^n(t))dt + \frac{1}{n}\sigma(\bar{X}^n(t^-))dN^{n\varphi}(t), \;\; \bar{X}_n(0) = x. \tag{3.11}$$

Consider the decomposition

$$\bar{X}^n(t) - x = \int_0^t b(\bar{X}^n(s))ds + \int_0^t \sigma(\bar{X}^n(s))\varphi^n(s)ds + \int_0^t \sigma(\bar{X}^n(s^-))(dN^{n\varphi^n}(s)/n - \varphi^n(s)ds). \tag{3.12}$$

We will show tightness for each term. Starting with the last; let's call it Q^n. The process $\{Q^n(t)\}_{t\geq 0}$ is a martingale with quadratic variation

$$[Q^n](t) \leq \frac{\|\sigma\|_\infty^2}{n^2} \mathbb{E} \int_0^1 \varphi^n(s)ds \leq \frac{\|\sigma\|_\infty^2}{n}(e + M) \tag{3.13}$$

where we've used the bound $b \leq e + \ell(b)$. By Burkholder-Gundy-Davis inequality,

$$\mathbb{E} \sup_{t\in[0,1]} |Q^n(t)| \leq c_1\, \mathbb{E}[Q^n](1)^{1/2} \to 0 \tag{3.14}$$

for some constant $c_1 < \infty$. Thus, by Chebyshev, Q_n converges weakly/in probability to zero; hence, tight.

Tightness of the first two terms will follow from estimates that show uniform equicontinuity. This is okay because the first two terms along produce sample paths in $C[0,1]$, for which Arzela-Ascoli is sufficient for (pre-)compactness. We've shown the ODE limit first to perhaps make this feel a bit better. Carrying out the estimates gives

$$\int_s^t b(\bar{X}^n(s))ds \leq \|b\|_\infty(t-s), \tag{3.15}$$

$$\int_s^t \sigma(\bar{X}^n(s))\varphi^n(s)ds \leq \int_s^t e^{c\|\sigma\|_\infty} + \frac{\ell(\varphi^n(s))}{c}ds \leq \left(e^{c\|\sigma\|_\infty} + \frac{M}{c}\right)(t-s), \tag{3.16}$$

and tightness follows from the bounds being uniform in n and ω.

Laplace upper bound We play the usual game. Take some $\delta > 0$, choose M accordingly and sequence of controls $\{\varphi^n\}_n \subset \mathcal{A}_{b,M}$ that are δ optimizers for the representation formula. Then, we choose a subsequence (denoted by n still) for which the pair $(\bar{X}^n, \varphi^n)$ converges in distribution. Then,

$$\liminf \frac{-1}{n} \log \mathbb{E}\, e^{-nF(X^n)} \geq \liminf \mathbb{E}\, F(\bar{X}^n) + \int_0^1 \ell(\varphi_n(s))ds - 2\delta \tag{3.17}$$

$$\geq \mathbb{E}\, F(\bar{X}) + \int_0^1 \ell(\varphi(s))ds - 2\delta \tag{3.18}$$

$$\overset{?}{\geq} \mathbb{E}\, F(\bar{X}) + I(\bar{X}) - 2\delta \tag{3.19}$$

$$\geq \inf_{\varphi \in D[0,1]} \{F(\psi) + I(\psi)\} - 2\delta \tag{3.20}$$

where $\bar{X}$ and φ denotes the corresponding limits (which we don't know yet!) and the second inequality is due to Fatou and lower semi-continuity of ℓ. Thus, it remains to show the unjustified inequality, which becomes clear as we find the weak limit. On an ω-by-ω basis, consider $\gamma_n \to \gamma$ in S_M and $\psi_n \to \psi$ uniformly, we claim that

$$\int_0^1 \sigma(\psi_n(s))\gamma_n(s)ds \to \int_0^t \sigma(\psi(s))\gamma(s)ds. \tag{3.21}$$

Indeed, using $b \leq e + \ell(b)$, we first get the estimate

$$\left| \int_0^1 (\sigma(\psi_n(s)) - \sigma(\psi(s)))\gamma_n(s)ds \right| \leq \sup_{s \in [0,1]} |\sigma(\psi_n(s)) - \sigma(\psi(s))| \int_0^1 \gamma_n(s)ds \tag{3.22}$$

$$\leq \sup_{s \in [0,1]} |\sigma(\psi_n(s)) - \sigma(\psi(s))|(e + M) \to 0. \tag{3.23}$$

Moreover, since $\nu^{\gamma_n} \to \nu^{\gamma}$ in the weak topology and the map $s \mapsto \sigma(\psi(s))$ on $[0, 1]$ is bounded and continuous, we get

$$\int_0^1 \sigma(\psi(s))(\gamma_n(s) - \gamma(s))ds \to 0. \tag{3.24}$$

Then, we've shown that for all ω,

$$\bar{X}(t) - x = \int_0^1 b(\bar{X}(s))ds + \int_0^1 \sigma(\bar{X}(s))\varphi(s)ds; \tag{3.25}$$

from which, the unjustified inequality follows as $\varphi(\omega) \in U_{\bar{X}}$.

Laplace lower bound Again, pick your favorite $\delta > 0$, and choose ψ^* such that

$$F(\psi^*) + I(\psi^*) \leq \inf_{\psi \in D[0,1]} \{F(\psi) + I(\psi)\} + \delta \tag{3.26}$$

and let $\varphi \in U_{\psi^*}$ such that $\int_0^1 \ell(\varphi(s))ds \leq I(\psi^*) + \delta := M$. As of now, it is now clear that $\varphi \in \mathcal{A}_{b,M}$, so we will parameterize and approximate

$$\varphi_q(t) = \left(\varphi(t) \vee \frac{1}{q}\right) \wedge q. \tag{3.27}$$

As $q \uparrow \infty$, we can see that limits are well-behaved (by sandwiching and monotone convergence):

$$\int_0^1 \ell(\varphi_q(s))ds \to \int_0^1 \ell(\varphi(s))ds, \;\; \psi_q^* = x + \int_0^{\cdot} b(\psi^*(s)ds + \int_0^{\cdot} \sigma(\psi^*(s))\varphi_q(s)ds \to \psi^*. \tag{3.28}$$

Thus, we have

$$\limsup \frac{-1}{n} \log \mathbb{E}\, e^{-nF(X^n)} \leq \limsup \mathbb{E}\, F(\bar{X}^n) + \int_0^1 \ell(\varphi_q(s))ds = F(\psi_q^*) + \int_0^1 \ell(\varphi_q(s))ds \tag{3.29}$$

Taking $q \to \infty$,

$$F(\psi^*) + \int_0^1 \ell(\psi(s))ds \leq F(\psi^*) + I(\psi^*) + \delta \leq \inf_{\psi \in D[0,1]} \{F(\psi) + I(\psi)\} + 2\delta. \tag{3.30}$$

Finally, taking $\delta \downarrow 0$ completes the proof. □

3.4 Small-noise limit of pure jump processes

Now, we turn our attention to pure-jump processes in $\mathbb{R}^d$ of the form

$$X^\epsilon(t) = x + \epsilon \sum_{k=1}^K \int_{[0,1]\times\mathbb{R}_+} \nu_k(X^\epsilon(s)) 1_{[0,\lambda_k(X^\epsilon(s))/\epsilon]}(y) N(ds, dy). \tag{3.31}$$

Alternatively, it is perhaps useful to consider the generator

$$\mathcal{L}^\epsilon f(x) = \frac{1}{\epsilon} \sum_{k=1}^K \lambda_k(x) \left(f(x + \epsilon \nu_k(x)) - f(x) \right). \tag{3.32}$$

The jump rate $\lambda_k : \mathbb{R}^d \to \mathbb{R}_+$ and jump dynamics $\nu_k : \mathbb{R}^d \to \mathbb{R}^d$ are assumed to be Lipschitz. Moreover, we assume that there exists a $c > 0$ such that $|\log \lambda_k(x)| \leq c$ for all $x \in \mathbb{R}^d$. Then, we're again interested establishing an LDP for the small-noise limit as $\epsilon \to 0$.

Theorem 3.5. *For $\xi \in C[0,1]$ with $\xi(0) = x$, we let U_ξ be the set of measurable maps (controls) $\varphi = \{\varphi_i\}_{i=1}^K$ such that*

$$\xi(t) = x + \sum_{k=1}^K \int_{[0,1]\times\mathbb{R}_+} \nu_k(\xi(s)) 1_{[0,\lambda_k(\xi(s))]}(y) \varphi_k(s,y) ds dy. \tag{3.33}$$

Then, the sequence $(X^\epsilon)_{\epsilon>0}$ satisfies a LDP with rate ϵ and rate function

3.5 Application: importance sampling for rare events

Importance sampling is a popular method for reducing variance for Monte Carlo estimators, which becomes particularly handy when we're trying to estimate probabilities of rare events. Let's fix a set $A \subset \mathbb{R}^d$ and we're interested in the probability that a random variable Y with law μ falls in A. The naive method is to sample $\{Y_j^n\}_{j=1}^n$ i.i.d. from μ and construct the estimator

$$\frac{1}{n} S^n = \frac{1}{n} \sum_{j=1}^n 1_{Y_j^n \in A}.$$

Let $p_n = \mathbb{P}(S^n/n \in A)$ and let's assume that Y has well-defined exponential moments such that the moment generating

function exists everywhere. Then, by Cramér's theorem, we know that $\mathbb{P} \circ (S_n/n)^{-1}$ satisfies the large deviation principle with rate L where

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^d} \langle \alpha, \beta \rangle - \log H(\alpha), \quad H(\alpha) = \log \mathbb{E}\, e^{\langle \alpha, Y \rangle}. \tag{3.34}$$

4 High-dimensional Geometry

4.1 Projection along random directions

In this section, let's take a geometric view on Cramér's theorem in terms of projections onto ℓ^2-balls. Consider a random vector $X^{(n)} = (X^{(1)}, \dots, X^{(n)}) \sim \mu^{\otimes n}$ and consider the "diagonal" direction on the ℓ^2-ball: $\iota^{(n)} = n^{-1/2}(1, 1, \dots) \in S^{n-1}$. Then, Cramér's theorem provides a large deviation principle for the (normalized) projection of the random vector X in the direction ι

$$\frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{\sqrt{n}} \langle \iota, X \rangle,$$

which has a rate function $I_\iota = \Lambda^*$. Perhaps it is natural to ask what happens when we choose some other (random) direction $\theta \in S^{n-1}$? The punchline here is that: we get an LDP with the same rate function for almost every θ and ι is not one of them!

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where X is defined on. Let σ_{n-1} be the rotationally-invariant measure on the sphere S^{n-1}. Let $\mathbb{S} = \prod_{n=1}^\infty S^{n-1}$ and the directions we choose $\theta = (\theta^{(1)}, \theta^{(2)}, \dots) \in \mathbb{S}$ will have law σ. We remark the running assumption throughout the subsection:

Assumption 4.1. *Define the coordinate map $\pi_n : \theta = (\theta^{(1)}, \theta^{(2)}, \dots) \in \mathbb{S} \mapsto \theta^{(n)}$, and let γ be a standard Gaussian measure on $\mathbb{R}$.*

1. *(Marginals are rotationally-invariant) For all $n \in \mathbb{N}$, $\sigma \circ \pi_n^{-1} = \sigma_{n-1}$.*
2. *(Finite fourth-moment) For all $t \in \mathbb{R}$, $\int \Lambda(tu)^4 \gamma(du) < \infty$.*

Remark 4.2. The first assumption subsumes the case of independent directions, i.e., $\sigma = \otimes_{n=1}^\infty \sigma_{n-1}$, but dependencies between $\theta^{(n)}$'s are allowed. Moreover, if μ admits a density f, then sub-exponential tail, i.e., for $|x| > c_1$, $f(x) \leq c_2 e^{c_3 |x|^p}$ for $1 < p < \infty$.

Universality of the rate function is in the following sense.

Theorem 4.3 (Theorem 2.4 of [4])**.** *For σ-a.e. $\theta \in \mathbb{S}$, the sequence $(W_\theta^{(n)})_{n \geq 0}$ where*

$$W_\theta^{(n)} = \frac{1}{\sqrt{n}} \langle \theta^{(n)}, X^{(n)} \rangle$$

satisfies an LDP with a convex rate function $I_\sigma = \Psi^*$ where

$$\Psi(t) = \int \Lambda(tu) \gamma(du). \tag{4.1}$$

Remark 4.4. The universality is perhaps less surprising if $\sigma = \otimes_{n=1}^\infty \sigma_{n-1}$. Since the random variable I_σ is in the tail-σ-algebra generated by the $\theta^{(n)}$'s, it must be trivial and take constant values σ-almost-everywhere. However, the universality holds even for very dependent sequences of $\theta^{(n)}$, so long as the marginals are rotationally-invariant.

The proof builds on the Gaussian characterization of surface measures and some manipulation for applying Gartner-Ellis theorem. We will go through a few technical lemmas.

Lemma 4.5. *Let $\mathbb{A} = \prod_{n=1}^\infty \mathbb{R}^n$ denote the space of triangular arrays and let $R : \mathbb{A} \to \mathbb{A}$ be the map such that for all $z = (z^{(1)}, z^{(2)}, \dots) \in \mathbb{A}$, we have the transformation*

$$z^{(n)} \mapsto z^{(n)} / \|z^{(n)}\|$$

for each row n. Let $\bar{\pi}_n : z \in \mathbb{A} \mapsto z^{(n)}$ be the coordinate map onto the n-th row. If there is a measure $\zeta \in \mathcal{P}(\mathbb{A})$ such that $\zeta \circ \bar{\pi}_n^{-1} = \gamma^{\otimes n}$, then $\sigma = \zeta \circ R^{-1}$ satisfies Item 1 of Assumption 4.1. Conversely, if σ satisfies Item 1 of Assumption 4.1, then there is a $\zeta \in \mathcal{P}(\mathbb{A})$ such that $\sigma = \zeta \circ R^{-1}$.

Proof. This is due to the fact that if $Z^{(n)} \sim \gamma^{\otimes n}$, then $Z/\|Z\| \sim \sigma_{n-1}$. □

The above lemma implies that for every σ satisfying the assumptions, we can find an equivalent ζ such that

$$\frac{1}{\sqrt{n}} \langle \theta^{(n)}, X^{(n)} \rangle \text{ with } \theta \sim \sigma \equiv \frac{\sqrt{n}}{\|z^{(n)}\|} \cdot \frac{1}{n} \sum_{i=1}^n z_i^{(n)} X_i \text{ with } z \sim \zeta. \tag{4.2}$$

Therefore, proving an LDP for $(W_\theta^{(n)})_n$ that holds for θ σ-a.e. is equivalent to proving an LDP for $\left(\frac{\sqrt{n}}{\|z^{(n)}\|} W_z^{(n)} \right)_n$ with

$$W_z^{(n)} = \frac{1}{n} \sum_{i=1}^n z_i^{(n)} X_i$$

for z ζ-a.e. It is perhaps reasonable to think that, by Gartner-Ellis, $(W_z^{(n)})_n$ satisfies an LDP. Therefore, we want to show that the factor $\sqrt{n}/\|z^{(n)}\|$ in front in unproblematic. Indeed, we have a reason to believe this is true as

$$\frac{\sqrt{n}}{\|z^{(n)}\|} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i^2\right)^{-1/2} \longrightarrow 1 \tag{4.3}$$

under ζ due to the strong law of large numbers. Rigorously, we do this by defining *exponential equivalences* between random variables and show that this does not interfere with LDPs.

Definition 4.6. We say that two random variables ξ and $\tilde{\xi}$ are exponentially equivalent if for all $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(|\xi_n - \tilde{\xi}_n| > \delta) = -\infty. \tag{4.4}$$

Lemma 4.7. *Let $(\xi_n)_n$ be a sequence of random variables satisfying an LDP with rate function I, and let $\tilde{\xi}_n = a_n\xi_n$ where $a_n \longrightarrow 1$ is a deterministic sequence. If I has convex level sets, then $(\xi_n)_n$ and $(\tilde{\xi}_n)_n$ are exponentially equivalent.*

Proof. For any ϵ, there is an N be such that $|1 - a_n| < \epsilon$ for $n \geq N$. This means, for any $\delta > 0$, $|\tilde{\xi}_n - \xi_n| \geq \delta$ only if

$$|\xi_n| \geq \frac{\delta}{1 - a_n} \geq \frac{\delta}{\epsilon}. \tag{4.5}$$

So, for any fixed $\delta > 0$,

$$\limsup \frac{1}{n} \log \mathbb{P}(|\tilde{\xi}_n - \xi_n| > \delta) \leq \limsup \frac{1}{n} \log \mathbb{P}\left(|\xi_n| \geq \frac{\delta}{\epsilon}\right) \leq -\inf_{|x|>\delta/\epsilon} I(x) = -\left(I\left(\frac{\delta}{\epsilon}\right) \wedge I\left(\frac{-\delta}{\epsilon}\right)\right). \tag{4.6}$$

Since I has compact level sets, it has a global minimizer $\bar{x}$. Choose ϵ such that $|\delta/\epsilon| \geq |\bar{x}|$, then I is non-decreasing (non-increasing) when $x > \delta/\epsilon$ $(x < -\delta/\epsilon)$ by level sets being convex and hence, coercive. Thus, letting $\epsilon \downarrow 0$, we have exponential equivalence. □

Lemma 4.8. *Let $(\xi_n)_n$ and $(\tilde{\xi}_n)_n$ be exponentially equivalent random variables. If $(\xi_n)_n$ satisfies an LDP with rate function I, then so does $(\tilde{\xi}_n)_n$.*

Proof. For the large deviation upper bound, let E be a closed set. Then, for any $\delta > 0$, we can write

$$\limsup \frac{1}{n} \log \mathbb{P}(\tilde{\xi} \in E) \leq \limsup \frac{1}{n} \log \Big(\mathbb{P}(\xi_n \in E + \delta) + \mathbb{P}(|\xi_n - \tilde{\xi}_n| > \delta)\Big) \tag{4.7}$$

where $E + \delta = \{x : d(x, E) < \delta\}$. Then, using the fact that sum is exponentially equivalent to maximum, we can write

$$\limsup \frac{1}{n} \log \Big(\mathbb{P}(\xi_n \in E + \delta) + \mathbb{P}(|\xi_n - \tilde{\xi}_n| > \delta)\Big) \leq -\inf_{x\in E+\delta} I(x) \vee -\infty \tag{4.8}$$

which yields the large deviation upper bound by taking $\delta \downarrow 0$.

On the other hand, it suffices to prove the lower bound for balls $B_{2\delta}(y)$ for some y in the state space, $\delta > 0$. Observe that

$$\begin{aligned} \liminf \frac{1}{n} \log \mathbb{P}(\tilde{\xi}_n \in B_{2\delta}(x)) &= \liminf \frac{1}{n} \log \mathbb{P}(\tilde{\xi}_n \in B_{\epsilon}(y)) \vee \limsup \frac{1}{n} \log \mathbb{P}(|\xi_n - \tilde{\xi}_n| > 2\delta) && (4.9) \\ &\geq \liminf \frac{1}{n} \log \Big(\mathbb{P}(\tilde{\xi}_n \in B_{2\delta}(y)) + \mathbb{P}(|\xi_n - \tilde{\xi}_n| > 2\delta)\Big). && (4.10) \end{aligned}$$

Notice that $\{\xi_n \in B_\delta(y)\} \subset \{\tilde{\xi}_n \in B_{2\delta}(y)\} \cup \{|\xi_n - \tilde{\xi}_n| > 2\delta\}$, so

$$\liminf \frac{1}{n} \log \Big(\mathbb{P}(\tilde{\xi}_n \in B_{2\delta}(y)) + \mathbb{P}(|\xi_n - \tilde{\xi}_n| > 2\delta)\Big) \geq \liminf \frac{1}{n} \log \mathbb{P}(\tilde{\xi}_n \in B_\delta) = -\inf_{x\in B_\delta(y)} I(x). \tag{4.11}$$

Approximating open sets as union of balls completes the proof. □

Proof of Theorem 4.3. By the previous lemmas, all that is left to prove is an LDP for the sequence $(W_z^{(n)})_n$ for z ζ-a.e. For now, fix a $z \in \mathbb{A}$ and we will apply Gartner-Ellis. Consider the limiting log-MGF:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp \left(ntW_z^{(n)} \right) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp \left(t \sum_{i=1}^{n} z_i^{(n)} X_i^{(n)} \right). \tag{4.12}$$

By independence of X_i's, we can rewrite the above into

$$\lim_{n\to\infty} \frac{1}{n} \log \prod_{i=1}^{n} \mathbb{E} \exp \left(tz_i^{(n)} X_i^{(n)} \right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Lambda(tz_i^{(n)}). \tag{4.13}$$

First, for a fixed t, the finite fourth-moment condition above gives a law of large numbers for triangular array and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Lambda(tz_i^{(n)}) = \int \Lambda(tu)\gamma(du) = \Psi(t) \tag{4.14}$$

ζ-almost surely. Moreover, since Λ is convex and finite everywhere (by assumption), the above limit stays convex and finite, hence, continuous in t. Therefore, it suffices for the convergence in n to hold for countably many t and extend by continuity. Thus, the set of measure zero remains measure zero and convergence to Ψ holds ζ-a.e.

Lastly, in order to get a true LDP from Gartner-Ellis, we need some regularity on Ψ so that there is an exposing hyperplane for every point in the state space. This amounts to showing differentiability of Ψ, which is implied by $t \mapsto \Lambda(tu)$ being differentiable if we can swap integration and limits.

For a fixed t, take some $\delta = [-1, 1]$. Then, by mean value theorem, we can bound the difference quotient

$$\left| \frac{\Lambda(u(t+\delta)) - \Lambda(ut)}{\delta} \right| \leq \sup_{\alpha \in [-1,1]} |\Lambda'(u(t+\alpha))u| \leq |\Lambda'(u(t-1))u| + |\Lambda'(u(t+1))u| \tag{4.15}$$

where the inequality is due to the monotonicity of Λ'. However, by convexity,

$$\Lambda(u(t-1)) - \Lambda(u(t-2)) \leq \Lambda'(u(t-1))u(t-1) \leq \Lambda(ut) - \Lambda(u(t-1)) \tag{4.16}$$

and similarly for $\Lambda'(u(t+1))u$. Thus, $|\Lambda'(u(t-1))u| + |\Lambda'(u(t+1))u|$ in integrable by integrability assumption on Λ and Ψ is differentiable by dominated convergence.

Therefore, by Gartner-Ellis, $(W_z^{(n)})_n$ satisfies an LDP with rate function Ψ^* for ζ-a.e. z; or alternatively, $(W_\theta^{(n)})_n$ satisfies an LDP with rate function Ψ^* for σ-a.e. θ. □

Remark 4.9. We can prove the LDP for $(W_z^{(n)})_n$ via weak convergence. We can get the variational characterization of the rate function

$$\Psi^*(x) = \inf_{\nu(\cdot|z)} \left\{ \int \mathcal{R}(\nu(\cdot|z) \| \mu)\gamma(dz) : \int zx\nu(dx|z)\gamma(dz) \right\}. \tag{4.17}$$

Now, we've established the universality for large deviation rate functions for weighted sums of i.i.d. random variables up to a set of measure-zero. Let's examine this null set more carefully; in particular, we want to compare I_σ with I_ι, the rate function obtained from Cramér's theorem, as empirical sums are typically the object of interest. For $\mu = \gamma$, we have

$$\int \Lambda(ut)\gamma(du) = \int \frac{(ut)^2}{2}\gamma(du) = \frac{t^2}{2} = \Lambda(t), \tag{4.18}$$

so $I_\sigma = I_\iota$ coincide. However, you can see that Gaussianity played a huge role here. The next theorem characterizes the atypicality of Cramér's theorem in terms of concavity when compared to the Gaussian.

Theorem 4.10 (Theorem 2.5 of [4]). *Assume that $\Lambda(t) = \Lambda(-t)$ for all t. Then,*

1. *If $\Lambda \circ \sqrt{\cdot}$ is concave, then $I_\sigma \geq I_\iota$.*
2. *If $\Lambda \circ \sqrt{\cdot}$ is convex, then $I_\sigma \leq I_\iota$.*

3. *If $\Lambda \circ \sqrt{\cdot}$ is concave or convex but not linear, then $I_\sigma = I_\iota$ only at zero.*

Proof. First, suppose $\Lambda \circ \sqrt{\cdot}$ is concave. Let $Z \sim \gamma$, then using the symmetry and Jensen, we get

$$\Psi(t) = \mathbb{E}\,\Lambda(tZ) = \mathbb{E}\,\Lambda((t^2Z^2)^{1/2}) \leq \Lambda(\mathbb{E}(t^2Z^2)^{1/2}) = \Lambda(t). \tag{4.19}$$

Hence, $\Psi^* \geq \Lambda^*$ everywhere as claimed. Notice that the above chain of computation yields equality if and only if $\Lambda \circ \sqrt{\cdot}$ is linear or the argument inside is degenerate, which happens if and only if $t = 0$. So, let $t_x = \operatorname{argmax}_t\{tx - \Lambda(t)\}$, then

$$\Psi^*(x) \geq t_x x - \Psi(t_x) \geq t_x x - \Lambda(t_x) = \Lambda^*(x) \tag{4.20}$$

with the second inequality turning an equality if and only if $t_x = 0$. However, $t_x = 0$ occurs only when $\partial_x \Lambda^*(x) = 0$, which by symmetry and strict convexity within its domain, occurs if and only $x = 0$.

The same analysis can be repeated for the case of $\Lambda \circ \sqrt{\cdot}$ is convex with reversed inequalities. □

Therefore, Cramér's theorem is indeed atypical! Unless μ is a standard Guassian, the universal rate function is not that of the empirical measure scaling. For the class of distributions with scale parameter $\alpha > 0$ and shape $\beta > 1$ where

$$\mu_{\alpha,\beta}(dx) = \frac{1}{2\alpha\Gamma(1+\beta^{-1})} e^{-(|x|/\alpha)^\beta} dx, \tag{4.21}$$

we can make direct comparison between I_σ and I_ι via studying the concavity compared to the Guassian distribution.

4.2 Spectral of random matrices: limit theory

4.3 Spectral of random matrices: large deviation

References

[1] Amarjit Budhiraja and Paul Dupuis, *Analysis and approximation of rare events*, Representations and Weak Convergence Methods, Series Probability Theory and Stochastic Modelling **94** (2019).

[2] Frank den Hollander, *Large deviations (fields institute monographs vol 14)(providence, ri: American mathematical society)* (2000).

[3] Paul Dupuis and Hui Wang, *Importance sampling, large deviations, and differential games*, Stochastics: An International Journal of Probability and Stochastic Processes **76** (2004), no. 6, 481–508.

[4] Nina Gantert, Steven Soojin Kim, and Kavita Ramanan, *Cramér's theorem is atypical*, Advances in the mathematical sciences: Research from the 2015 association for women in mathematics symposium, 2016, pp. 253–270.

A Reading Notes

A.1 Week 1

1. Why is it about moment generating functions?

2. For the Legendre transform, at least in our case, we have the following relationship: $(f)^{*'}(p) = (f')^{-1}(p)$. Note that the -1 is the inverse of the function.

3. Why, intuitively, the answer is the Legendre transform of log mgf?

4. The rate of the tail $\mathbb{P}(S_n \geq 0)$ is estimated, and this is non-trivial since you cannot get it from CLT directly (you are not evaluating the tail at a fixed interval, the interval itself is shrinking)

5. how does page 11 work rigorously?

6. I don't see the open set is needed on page 15

7. Does the paradigm "rare events happen in the most likely of unlikely ways" say anything about the distribution conditioned on a rare event?

8. Chatterjee Dembo?

A.2 Week 2

1. Page 21: $\bar{\mu}_i^n$ is a random measure, so what does it mean to have $\mu_i^* = \mu^*$?

2. Relative entropy always well-defined? How to derive "from first principle" variational formulas, e.g., Donsker-Varadhan?

3. Hints on Problem 4 and 12? Precise definition of liminf?

A.3 Week 3

B Exercises