Representations and weak convergence methods for the analysis and approximation of rare events

A Short Course of the Scuola di Dottorato in Scienze matematiche, Università degli Studi di Padova

Paul Dupuis
Division of Applied Mathematics
Brown University
Providence, RI 02896
USA

Outline:

1. Introduction and Examples
2. General Theory and Relative Entropy
3. Canonical Problem I – Sanov's Theorem
4. Canonical Problem II – Small Noise Diffusions
5. Freidlin-Wentsell Theory and Moderate Deviations
6. Processes That Are Not Functionals of an IID Noise Process
7. Extracting Information From the Variational Problem
8. An Overview of Importance Sampling for Rare Events
9. The Subsolutions Approach to Importance Sampling
10. The Subsolutions Approach to Importance Sampling, Cont'd
11. The Empirical Measure of a Markov Chain
12. Current Developments and Related Problems

The goal of these notes is to introduce the reader to methods for characterizing and analyzing rare events, and for the construction and analysis of related Monte Carlo numerical approximations. The approach to both topics is based on weak convergence theory and relative entropy representations for exponential integrals. Some of the ideas and methods presented in these notes first appeared in *A Weak Convergence Approach to the Theory of Large Deviations* with Richard Ellis, which in particular has a very detailed discussion of many of the nice properties of relative entropy we will use. Many new topics, including infinite dimensional problems and the analysis of Monte Carlo, will appear in a forthcoming book with Amarjit Budhiraja with the same title as these notes: *Representations and Weak Convergence Methods for the Analysis and Numerical Approximation of Rare Events.*

These notes were prepared as part of a short course given at Dipartimento di Matematica, Università degli Studi di Padova, from 20-31 May, 2013. The author would like to thank the department, and in particular Markus Fischer, for their warm hospitality.

Lecture 1: Introduction and Examples

1 The setting and statement of a large deviation principle

Let S denote a Polish space with Borel σ-algebra $\mathcal{B}(S)$. Typical examples of S in these notes will be $\mathbb{R}^d$ (Euclidean d-dimensional space), $C\left([0,T]:\bar{S}\right)$ (the set of continuous functions mapping $[0,T]$ to $\bar{S}$) and $\mathcal{P}(\bar{S})$ (the set of probability measures on $\left(\bar{S},\mathcal{B}(\bar{S})\right)$), where $\bar{S}$ is itself a Polish space.

Let $\{X_n, n\in\mathbb{N}\}$ be S-valued random variables on $(\Omega,\mathcal{F},P)$, with distributions

$$\mu_n(B) = P\{X_n\in B\}, \quad B\in\mathcal{B}(S).$$

Definition 1 A function $I: S\to[0,\infty]$ is called a *rate function* if $\{x\in S: I(x)\le M\}$ is compact for all $M\in[0,\infty)$.

Definition 2 The sequence of random variables $\{X_n, n\in\mathbb{N}\}$ (or equivalently the sequence of distributions $\{\mu_n, n\in\mathbb{N}\}$) is said to satisfy the *large deviation principle (LDP)* with rate I, if I is a rate function, and if

1. for all open sets $O\in\mathcal{B}(S)$

$$\liminf_{n\to\infty}\frac{1}{n}\log P\{X_n\in O\}\ge -\inf_{x\in O} I(x),$$

2. for all closed sets $F\in\mathcal{B}(S)$

$$\limsup_{n\to\infty}\frac{1}{n}\log P\{X_n\in F\}\le -\inf_{x\in F} I(x).$$

In a *very* rough sense, one can think of this as saying

$$P\{X_n\in B_\delta(x)\}\approx e^{-nI(x)},$$

where $\delta>0$ is small and $B_\delta(x)=\{y: d(y,x)<\delta\}$, with d the metric on S. For $C\in\mathcal{B}(S)$ let $I(C)=\inf_{x\in C} I(x)$. If $I(C^\circ)=I(\bar{C})=I(C)$, then C is called an I-continuity set and we have

$$\lim_{n\to\infty}\frac{1}{n}\log P\{X_n\in C\}=I(C).$$

Variational problems arise naturally in large deviation problems because of the following elementary consequence of exponential scaling.

Lemma 1 *Let sequences* $\{a_n\}, \{b_n\} \subset [0,\infty]$ *be given such that*

$$-\frac{1}{n}\log a_n \to u \in [0,\infty]$$
$$-\frac{1}{n}\log b_n \to v \in [0,\infty].$$

Then

$$-\frac{1}{n}\log\left(a_n + b_n\right) \to \min\left\{u, v\right\}.$$

Thus if a_n and b_n are probabilities scaling like $a_n \approx e^{-nu}$ and $b_n \approx e^{-nv}$, then the decay rate of $a_n + b_n$ is given by the smaller of u and v.

Remark 1 The scaling parameter $n \in \mathbb{N}$ is sometimes replaced by $\varepsilon > 0$, with n corresponding to $1/\varepsilon$.

2 Examples

The following examples illustrate various applications. Proofs of the LDP for some (but not all) of these models will be given. Some will also be used in the discussion of Monte Carlo methods.

Example 1 (Multi-dimensional random walk, insurance risk) Let Z_i be independent and identically distributed (iid) with distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$. For $x \in \mathbb{R}^d$ and $n \in \mathbb{N}$ define

$$X_{i+1}^n = X_i^n + \frac{1}{n}Z_i, \quad X_0^n = x,$$

and the piecewise linear interpolation

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right].$$

Assume that $EZ_i < 0$ component-wise and for $M \in (0,\infty)^d$ let

$$\tau^n = \inf\left\{t \geq 0 : X^n(t)_j \geq M_j \text{ for some } j = 1, \ldots, d\right\}.$$

The problem of estimating $P\left\{\tau^n < \infty\right\}$ arises in insurance risk, with the correlation between different components of Z_i modeling correlation between different sectors or firms.[1]

Assume the log-moment generating function satisfies

$$H(\alpha) \doteq \log Ee^{\langle \alpha, Z_i\rangle} < \infty \text{ for all } \alpha \in \mathbb{R}^d.$$

[1] In fact, the origins of large deviation theory can be traced to applications in insurance though H. Cramér.

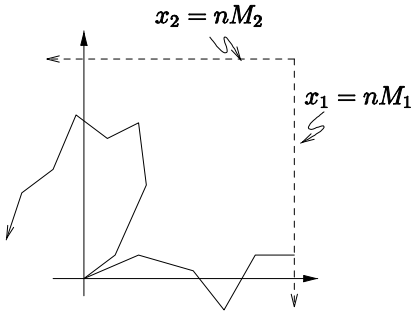

Figure 1: Two dimensional escape set, unscaled random walk

Let $L(\beta)$ be the Legendre transform of H:

$$L(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H(\alpha) \right], \quad \beta \in \mathbb{R}^d.$$

Then for each $T < \infty$, $\{X^n, n \in \mathbb{N}\}$ satisfies the LDP on $C\left([0,T] : \mathbb{R}^d\right)$ with rate function

$$I_T(\phi) = \begin{cases} \int_0^T L(\dot{\phi}(t))dt & \text{if } \phi \text{ is absolutely continuous and } \phi(0) = x, \\ \infty & \text{otherwise.} \end{cases}$$

Thus for a given ϕ, the rough interpretation gives an estimate for the probability that X^n "tracks" ϕ in the form

$$P\left\{ \sup_{0 \le t \le T} |X_n(t) - \phi(t)| \le \delta \right\} \approx e^{-nI_T(\phi)}.$$

Although it does not follow directly from the LDP, one can use the large deviation estimates on $[0,T]$ (or argue directly using weak convergence methods) that

$$\begin{aligned} &-\frac{1}{n} \log P\left\{\tau^n < \infty\right\} \\ &\quad \to \quad \inf\left\{I_T(\phi) : \phi(T)_j \ge M_j \text{ for some } j = 1, \ldots, d,\, T < \infty\right\}. \end{aligned} \tag{1}$$

The idea behind the reduction to finite time estimates is straightforward. One first shows using the upper bound alone that if the event is to occur at all, then it must happen with overwhelming probability before some fixed finite time T (see, e.g., [34, Lemma 2.2, Chapter 5]). Specifically, one shows that given any $K < \infty$ there is $T < \infty$ such that

$$\limsup_{n \to \infty} -\frac{1}{n} \log P\left\{\tau^n \in [T, \infty)\right\} \le K.$$

It follows that $P\left\{\tau^n < \infty\right\}$ and $P\left\{\tau^n \in [0,T]\right\}$ have the same decay for sufficiently large but finite T, and an application of the LDP to $P\left\{\tau^n \in [0,T]\right\}$ gives (1).

Remark 2 While the large deviation approximation in this and other examples is guaranteed to give the correct rate of decay, depending on the particular application, the estimate itself may not be as accurate as one needs. The rate of decay is often well suited to qualitative issues (e.g., control and design). If a more accurate approximation to $P\{\tau^n < \infty\}$ is desired then the large deviation information is very useful in the design of Monte Carlo schemes. This application is the subject of Lectures 8–10.

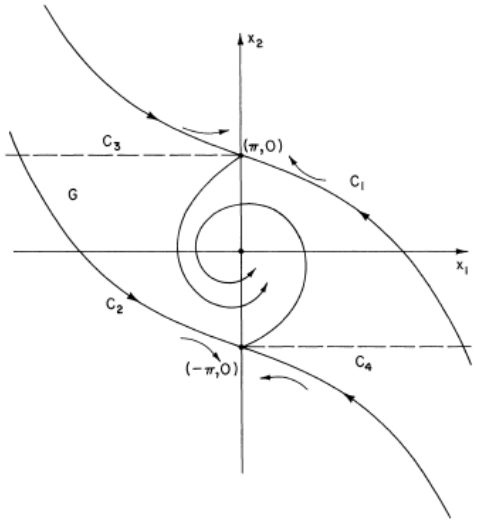

Figure 2: Dynamics of tracking loop with no noise

Example 2 (Metastability for diffusion processes, a PLL type example). Various algorithms in adaptive control, suboptimal filtering, and elsewhere are designed to reject noise and keep a parameter near a desired operating point [44, 42, 19]. Large deviation theory gives natural measures of the performance of these algorithms. An example is the following diffusion model of a phase-locked loop:

$$\begin{aligned} dX_1^\varepsilon &= -a\pi X_1^\varepsilon dt + b\left(\sin \pi X_2^\varepsilon dt + \sqrt{\varepsilon} dW\right) \\ dX_2^\varepsilon &= -\pi X_1^\varepsilon dt \end{aligned}$$

with $-X_2^\varepsilon$ a measure of the "tracking error." Here a and b are parameters to be selected for the loop design, and higher order loops have a larger dimension and more parameters to select. The "noiseless" system ($\varepsilon = 0$) is illustrated in Figure 2. The diffusion model arises from a device driven by wide bandwidth noise and high carrier frequency ω^γ as indicated in Figure 3, after the noise is approximated by a Brownian motion and high frequency

terms due to the double angle formula and the multiplexer are dropped, with $X_2^\varepsilon = \theta - \hat{\theta}^\rho$.

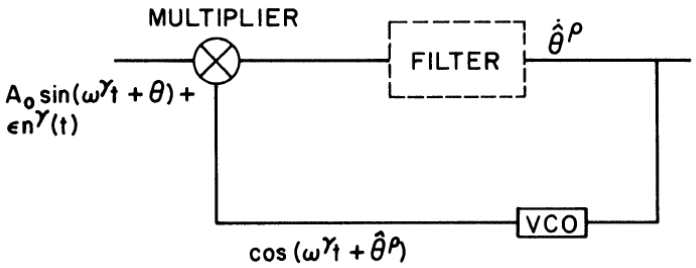

Figure 3: Tracking loop driven by signal plus noise

Performance measures would include, e.g., $P_0\{\tau^\varepsilon < T\}$, where $\tau^\varepsilon = \inf\{t \geq 0 : |X_2^\varepsilon(t)| \geq \pi\}$, and P_0 denotes probability given $X^\varepsilon(0) = 0$. Given ϕ with $\phi(0) = 0$, let $\mathcal{S}_\phi$ be the set of $u \in L^2([0,T] : \mathbb{R})$ such that for all $t \in [0,T]$

$$\begin{aligned} \phi_1(t) &= -\int_0^t a\pi\phi_1(s)ds + \int_0^t b\left(\sin \pi\phi_2(s) + u(s)\right) ds \\ \phi_2(t) &= -\int_0^t \pi\phi_1(s)ds \end{aligned}$$

The rate function for $\{X^\varepsilon, \varepsilon \in (0,1)\}$ with this initial condition is

$$I_T(\phi) = \inf\left\{\int_0^T \frac{1}{2}u(t)^2 dt : u \in \mathcal{S}_\phi\right\},$$

where the infimum over the empty set is ∞. (We call this the *control form* of the rate function since we view u as a control and ϕ as a controlled state. In contrast, the rate function for the previous example was in a *calculus of variations* form). It follows from the LDP that

$$-\varepsilon \log P_0\{\tau^\varepsilon < T\} \to \inf\{I_T(\phi) : |\phi_2(t)| \geq \pi \text{ for some } t \in [0,T]\}.$$

Again, non-asymptotic approximations to $P_0\{\tau^\varepsilon < T\}$ are very useful.

Example 3 (Empirical measure large deviations and MCMC) Consider an ergodic Markov chain $\{X_i, i \in \mathbb{N}_0\}$ with state space S and unique stationary distribution π. The *empirical measure* or *normalized occupation measure* of the chain is defined for $n \in \mathbb{N}$ by

$$\mu^n(A) = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{X_i}(A), \quad A \in \mathcal{B}(\mathcal{S})$$

where δ_x is the Dirac measure that puts probability 1 at x.

Under appropriate conditions, μ^n converges in an appropriate topology to π with probability one (w.p.1). In fact, this property is used in many important applications in the physical and biological sciences, statistics, and elsewhere as a method of numerically approximating π. In this setting, a large deviation principle for $\{\mu^n, n \in \mathbb{N}\}$ will give information on the likelihood that μ^n is near an alternative "target" measure besides π. Since many chains can have the same invariant distribution, the rate function can then be used to compare the numerical efficiency of the possible chains. Large deviation theory for the empirical measure is also used in many other areas, such as information theory and statistics. We will discuss this example and such an application in Lecture 11. The large deviation theory for the empirical measure of a Markov chain was originally developed in the papers [12, 13].

Example 4 (Queueing and data loss) An area where large deviation has been very active is in the analysis of stochastic network models, and especially models for communication. In this example we present a simple model that involves choosing parameters to achieve a desired loss rate.

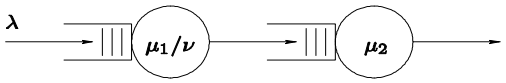

Figure 4: A tandem queue with two service rates

The tandem queuing model is depicted in Figure 4. The second queue has a finite buffer, and data is lost when the process reaches this buffer. When the second queue is small the first queue serves at rate μ_1. However, when the second queue exceeds a threshold the first queue reduces its service rate to ν. The problem is to determine a threshold so that a prescribed and very small loss rate holds.

Since the buffer is expected to be relatively large it is scaled by n, as is the threshold, which is given by θn with $\theta \in (0, 1)$. The dynamics and partition of the state space for the system as well as the state space for the rescaled system (introduced below) are illustrated in Figure 5. The queueing process $(Q_1(t), Q_2(t))$ is modeled as a jump Markov model with the indicated rates. Since a queue cannot be negative, jump rates are zero for jumps that would cause the state to leave $(\mathbb{N}_0)^2$. The problem of interest is to estimate quantities such as

$$P_{(1,0)} \{Q_2 \text{ exceeds } n \text{ before } (Q_1, Q_2) \text{ reaches } (0, 0)\}$$

where $P_{(1,0)}$ denotes probability given $(Q_1(0), Q_2(0)) = (1, 0)$.

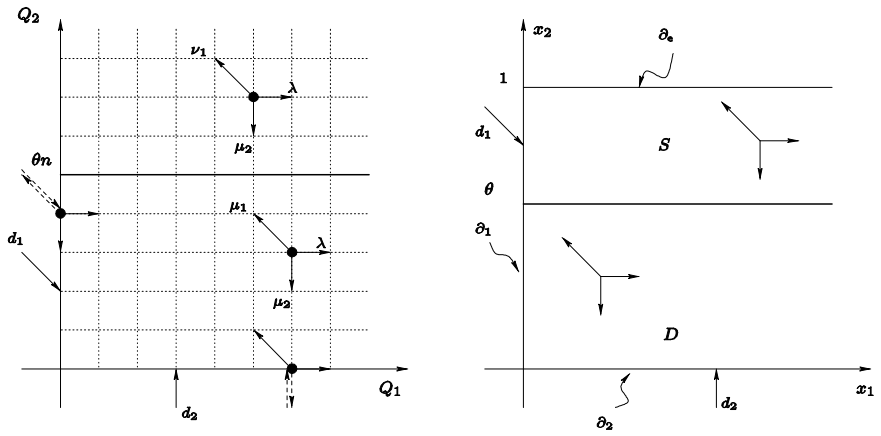

Figure 5: Jump rates and partition of the state space for the scaled system.

Large deviation estimates can be proved for the scaled system defined by $(Q_1^n(t), Q_2^n(t)) = (Q_1(nt), Q_2(nt))/n$ [21]. Expressing the quantity mentioned above in terms of the scaled system gives

$$P_{(1/n,0)}\left\{Q_2^n \text{ exceeds 1 before } (Q_1^n, Q_2^n) \text{ reaches } (0,0)\right\}.$$

Although this quantity involves an a priori potentially unbounded time interval, one can show as with the first example that if the event is to occur at all, it will happen with overwhelming probability before some fixed time T, which allows a reduction to the finite time LDP.

Owing to the presence of boundaries and an interface across which the rates suffer a discontinuity, the large deviations analysis presents a number of difficulties, and this example falls into the category of large deviation theory for processes with "discontinuous statistics" [15, 16, 45].

Example 5 (Occupancy models) Consider a large number of urns into which a large number of tokens will be distributed according to some randomized rule. A quantity of interest in this context is the empirical measure according to the number of tokens. Thus if there are n urns and Tn tokens to be distributed, then we are interested in the distribution of

$$\Gamma^n(T) \doteq \left(\Gamma_0^n(T), \Gamma_1^n(T), \ldots, \Gamma_J^n(T), \Gamma_{J+}^n(T)\right),$$

where $\Gamma_j^n(T)$ is the fraction of urns that contain exactly j tokens, and $\Gamma_{J+}^n(T)$ is the faction that contain strictly more than J tokens, after all have been distributed. (One can also consider the infinite dimensional empirical measure, but to simplify we restrict to the finite dimensional case here.)

To be specific, suppose that the urn chosen for a given token is selected uniformly, and independent of the selection for all other tokens. One can

then consider the evolution of the occupancy vector $\Gamma^n(i/n)$, where at (continuous) time i/n exactly i tokens have been placed. The placement of the next token into the various categories indexed by $j = 0, 1, \ldots, J, J+$ will be determined by the vector $\Gamma^n(i/n)$, since each urn is equally likely to receive the next token. Let y_i^n have the conditional distribution

$$P\left\{ y_i^n = j \middle| \mathcal{F}_i^n \right\} = \Gamma_j^n(i/n),$$

where $\mathcal{F}_i^n = \sigma\left(\Gamma^n(k/n), 0 \le k \le i\right)$. Since when $y_i^n = j$ the class of urns of type j is reduced by 1 while the class of type $j+1$ is increased by 1, the dynamics of $\Gamma^n(i/n)$ are given by

$$\Gamma^n((i+1)/n) = \begin{cases} \Gamma^n(i/n) + \frac{1}{n}\left(e_{y_i^n+1} - e_{y_i^n}\right) & \text{if } j \in \{0, 1, \ldots, J\} \\ \Gamma^n(i/n) & \text{if } j = J+ \end{cases}.$$

This is the same scaling as we have seen in some of the other examples, and indeed one can prove an LDP for the piecewise linear continuous time process defined by

$$\Gamma^n(t) = \Gamma^n(i/n) + \left[\Gamma^n((i+1)/n) - \Gamma^n(i/n)\right](nt - i), \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right].$$

This particular occupancy problem has a number of applications for which a large deviations analysis is relevant. One example is to the testing of random number generators [32], where the urns correspond to a finite uniform partition of $[0, 1]$, the tokens to $U[0, 1]$ iid random variables, and a token is assigned to an urn if the random variable falls into the corresponding subset of the partition. The distribution of $\Gamma^n(T)$ gives a very sensitive measure of the degree to which the variates are truly iid $U[0, 1]$. Another application is to the dimensioning of optical switches in communication networks [47]. A last application is to the empirical distribution of the number of lottery players who have selected the same combination (note that in many lotteries the number of combinations and players may be on the order of 10^8).

Other rules of placement that increase or decrease the likelihood that a given urn is selected depending on its current state are of interest, and the various schemes go by names such as Bose–Einstein, Maxwell–Boltzmann, and Fermi–Dirac statistics [48].

Example 6 (Performance analysis in rare event Monte Carlo) Our final example concerns the analysis of Monte Carlo schemes (such as importance sampling) which might be used to provide approximations more accurate than the large deviation approximation. To be concrete, we consider importance sampling, and its application to the first example. The standard measure of accuracy for such a scheme is the variance of a single sample, and due to unbiasedness the minimization of variance is equivalent

to minimization of the second moment. As we discuss in Lecture 9, this second moment ends up being characterized as an exponential integral, and specifically one of the form

$$E\left[1_{\{\tau^n<\infty\}}\prod_{i=0}^{n\tau^n-1}e^{-\left\langle\alpha_i^n(X_i^n),Z_i\right\rangle+H(\alpha_i^n(X_i^n))}\right],$$

where τ^n, X_i^n, Z_i, and H are all defined as in the example, and $\alpha_i^n(\cdot)$ characterizes the specific importance sampling scheme used. As we will see in the Lecture 10, the same weak convergence methods used to analyze the original problem can also be used to evaluate these integrals, and thereby allow for the design and comparison of efficient Monte Carlo schemes.

In closing this lecture we should mention that these process models and applications are just a small fraction of the classes which have been studied and which one would like to study. Among the many possible generalizations we could mention, we limit ourselves here to observing that the large deviation analysis of infinite dimensional models has become very active in recent years. Also, much more complex noise models have been studied. An elementary example in that direction would be to replace the iid structure of Example 1 by an ergodic Markov process. The study of other functionals besides the escape probability functionals emphasized here occurs in many applications. Finally, we mention that although we have not discussed them there are many applications in mathematical finance, and among other sources the reader can consult Glasserman's book [35].

Lecture 2: General Theory and Relative Entropy

1 General Theory

1.1 Laplace formulation

The definition of an LDP is phrased in terms of lower and upper bounds for open and closed sets. However, as with weak convergence it can be simpler in many ways to use a formulation in terms of expected values of bounded continuous functions. In the setting of a Polish space, we have the following result [14, Theorem 1.2.3].

Theorem 1 *A sequence of random variables $\{Y_n\}$ that takes values in a Polish space S satisfies the LDP with rate function I if and only if it satisfies the following Laplace Principle with rate I:*

- *I is a rate function (i.e., the set $\{x : I(x) \leq M\}$ is compact for any $M < \infty$), and*
- *for any bounded and continuous function $F : S \to \mathbb{R}$,*

$$\lim_{n\to\infty} -\frac{1}{n} \log E e^{-nF(Y_n)} = \inf_{x\in S} \left[I(x) + F(x)\right].$$

The proof of this fact is very similar to one of the standard proofs of the Portmanteau Theorem, and involves showing how to properly approximate such an f by combinations of indicator functions of sets, and conversely. The following facts are easy to verify:

- rate functions are unique, and
- a rate function attains its infimum over any closed set.

Note that it is always true that $I(S) = 0$, and so in particular there is always at least one point x^* where $I(x^*) = 0$.

Large deviations gives us not only quantitative information (how likely is the event $\{Y_n \in C\}$), but also qualitative information (given that the unlikely event $\{Y_n \in C\}$ occurred, how did it happen).

Theorem 2 *Assume that C is closed and that $I(C) = I(C^\circ) < \infty$. Let $G = \{x : I(x) = I(C)\}$, and $G^\varepsilon = \{x : d(x, G) < \varepsilon\}$. If $\{Y_n\}$ satisfies the LDP with rate I, then for any $\varepsilon > 0$*

$$P\left\{Y_n \in G^\varepsilon \,|Y_n \in C\right\} \to 1.$$

Proof. We claim that there is $\delta > 0$ such that $I(C \backslash G^\varepsilon) - I(C) = \delta$. If not, one can find $x_i \in C \backslash G^\varepsilon$ such that $I(x_i) \leq I(C) + 1/i$. Since level sets of I are compact, there is a subsequence x_{i_k} that converges to $x^* \in C \backslash G^\varepsilon$ with $I(x^*) = I(C)$, which contradicts the definition of G^ε. By Bayes' rule

$$\begin{aligned} P\left\{Y_n \in G^\varepsilon \cap C \,|Y_n \in C\right\} &= 1 - P\left\{Y_n \in C \backslash G^\varepsilon \,|Y_n \in C\right\} \\ &= 1 - \frac{P\left\{Y_n \in C \backslash G^\varepsilon\right\}}{P\left\{Y_n \in C\right\}}. \end{aligned}$$

Since

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left\{Y_n \in C\right\} \geq -I(C^\circ) = -I(C)$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left\{Y_n \in C \backslash G^\varepsilon\right\} \leq -I(C) - \delta,$$

the result follows. □

1.2 Contraction principle

Another parallel with weak convergence theory is the following analogue of the Continuous Mapping Theorem.

Theorem 3 (Contraction Principle) *Let $\{Y_n, n \in \mathbb{N}\}$ be S_1-valued random variables that satisfy the LDP with rate function I. Let $G : S_1 \to S_2$, where S_2 is another Polish space and G is continuous. Then $\{G(Y_n), n \in \mathbb{N}\}$ satisfy the LDP with rate function*

$$J(y) = \inf\left\{I(x) : G(x) = y\right\}.$$

Proof. There are only two items to prove. The first is that $\{y \in S_2 : J(y) \leq K\}$ is compact. However, this is automatic, since this set is just the forward image of the compact set $\{x \in S_1 : I(x) \leq K\}$ under the continuous function G. The second is that for any bounded and continuous function $F : S_2 \to \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n} \log E e^{-nF(G(Y_n))} = \inf_{y\in S_2} \left[F(y) + J(y)\right].$$

This is also easy, since

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n} \log E e^{-nF(G(Y_n))} &= \inf_{x\in S_1} \left[F(G(x)) + I(x)\right] \\ &= \inf_{y\in S_2} \inf_{x\in S_1 : G(x)=y} \left[F(G(x)) + I(x)\right] \\ &= \inf_{y\in S_2} \left[F(y) + J(y)\right]. \end{aligned}$$

□

2 Relative Entropy

2.1 Definition and elementary properties

A key ingredient in the weak convergence approach to analyzing large deviations is the famous *relative entropy function*. Let S be a Polish space, and consider probability measures μ and θ an S. We define $R(\mu\,\|\theta)$, the relative entropy of μ given θ, by

$$\int_S \left(\frac{d\mu}{d\theta}\right) \log\left(\frac{d\mu}{d\theta}\right) d\theta = \int_S \log\left(\frac{d\mu}{d\theta}\right) d\mu$$

if $\mu \ll \theta$, and set $R(\mu\,\|\theta) = \infty$ otherwise. In the definition the convention $0 \log 0 = 0$ is used. Since $x \log x$ is bounded from below for $x \geq 0$, the first integral is always well defined if $\mu \ll \theta$. Relative entropy places a central role in information theory, statistical mechanics, and other disciplines, and is a well studied quantity. The second item in the following lemma follows from the *Donsker-Varadhan formula for relative entropy*. Let $C_b(S)$ denote the bounded and continuous functions from S to $\mathbb{R}$. Then for any pair $\mu, \theta \in \mathcal{P}(S)$,

$$R(\mu\,\|\theta) = \sup_{g \in C_b(S)} \left[\int_S g d\mu - \log \int_S e^g d\theta\right].$$

For a proof see [14, Section C.2].

Lemma 4 (Elementary Properties of Relative Entropy)

- $R(\mu\,\|\theta) \geq 0$, *and* $R(\mu\,\|\theta) = 0$ *if and only if* $\mu = \theta$.
- $R(\mu\,\|\theta)$ *is a convex function and lower semicontinuous function of* $(\mu, \theta) \in \mathcal{P}(S)^2$.

Proof. In proving the first item we can assume $R(\mu\,\|\theta) < \infty$. In this case $d\mu/d\theta$ is well defined. We use that $s \log s \geq s - 1$ with equality if and only if $s = 1$. Thus

$$R(\mu\,\|\theta) = \int_S \frac{d\mu}{d\theta}\left(\log \frac{d\mu}{d\theta}\right) d\theta \geq \int_S \left(\frac{d\mu}{d\theta} - 1\right) d\theta = 0,$$

and equality holds only when $d\mu/d\theta = 1$, which requires $\mu = \theta$. □

2.2 Variational formula — a representation for exponential integrals

Large deviation theory, as phrased in terms of a Laplace principle, amounts to calculating the asymptotics of certain scaled exponential integrals, and expresses the asymptotics in terms of a variational problem. Hence it should come as no surprise that a variational representation for exponential integrals would be useful in large deviation analysis.

Lemma 5 (Representation for exponential integrals) *Given any bounded and measurable $f : S \to \mathbb{R}$,*

$$-\log \int_S e^{-f} d\theta = \inf_{\mu \in \mathcal{P}(S)} \left[\int_S f d\mu + R(\mu \,\|\theta) \right].$$

Proof. It suffices to prove that

$$-\log \int_S e^{-f} d\theta = \inf \left[\int_S f d\mu + R(\mu \,\|\theta) : R(\mu \,\|\theta) < \infty \right].$$

One can formally guess by a Lagrange multiplier argument that the minimizer should be given by

$$\frac{d\mu^*}{d\theta}(x) = e^{-f(x)} \cdot \frac{1}{\int_S e^{-f} d\theta}.$$

Since under $R(\mu \,\|\theta) < \infty$ μ is absolutely continuous with respect to θ, and since θ is absolutely continuous with respect to μ^*, it follows that μ is absolutely continuous with respect to μ^*. Using the definition of relative entropy twice, we write

$$\begin{aligned} \int_S f d\mu + R(\mu \,\|\theta) &= \int_S f d\mu + \int_S \log \left(\frac{d\mu}{d\theta} \right) d\mu \\ &= \int_S f d\mu + \int_S \log \left(\frac{d\mu}{d\mu^*} \right) d\mu + \int_S \log \left(\frac{d\mu^*}{d\theta} \right) d\mu \\ &= -\log \int_S e^{-f} d\theta + R(\mu \,\|\mu^*). \end{aligned}$$

Now use that $R(\mu \,\|\mu^*) \geq 0$ with equality only when $\mu = \mu^*$. This not only proves the formula, but incidentally identifies the minimizer. □

Remark 1 The assumption that f is bounded can be weakened. For example, it holds if f is uniformly bounded from below, which allows one to represent probabilities, i.e., $-\log \theta(A)$. Under additional properties of θ, such as a finite moment generating function when $S = \mathbb{R}^d$, one can extend to f that grow no faster than linearly. This will be useful later in the notes.

2.3 Chain rule

To make the representation for exponential integrals useful, one needs to decompose relative entropy for complex measures into relative entropies with respect to more basic units. The chain rule does exactly that, and will be used frequently. For a proof, see [14, Theorem B.2.1].

Lemma 6 (Chain rule) *Suppose that S is of product form, $S = S_1 \times S_2$, where both S_1 and S_2 are Polish. If $(\mu, \theta) \in \mathcal{P}(S)^2$, and if each distribution is*

factored into its marginal distribution on S_1 times a conditional distribution on S_2 given S_1:

$$\mu(dx_1 \times dx_2) = [\mu]_1(dx_1)[\mu]_{2|1}(dx_2\,|x_1\,),$$

$$\theta(dx_1 \times dx_2) = [\theta]_1(dx_1)[\theta]_{2|1}(dx_2\,|x_1\,),$$

then

$$R\,(\mu\,\|\theta\,) = R\,([\mu]_1\,\|[\theta]_1\,) + \int_{S_1} R\left([\mu]_{2|1}(\cdot\,|x_1\,)\,\middle\|[\theta]_{2|1}(\cdot\,|x_1\,)\right)[\mu]_1(dx_1).$$

2.4 Control representations for structured measures

Owing to the role it plays in the representations, we will refer to the measure appearing in the second position in relative entropy, i.e., θ in $R\,(\mu\,\|\theta\,)$, as the "base" measure. When the base measure is structured, such as when θ is a product measure or a Markov measure, a more useful, control-theoretic representation can be found in terms of component measures that make up θ.

Here is an example. Suppose that the random variables (X_1, X_2) have joint distribution θ and $(\bar{X}_1, \bar{X}_2)$ have distribution μ, on some probability space $(\Omega, \mathcal{F}, P)$. Further suppose that the θ measure corresponds to *independent* random variables, so that $[\theta]_{2|1}(\cdot\,|x_1\,) = [\theta]_2(\cdot\,)$ for some probability measure $[\theta]_2$. If $f : S_1 \times S_2 \to \mathbb{R}$ is bounded and measurable, then

$$-\log Ee^{-f(X_1,X_2)} = \inf_{\mu\in\mathcal{P}(S)} E\left[f(\bar{X}_1, \bar{X}_2) + \sum_{i=1}^{2} R\,(\bar{\mu}_i\,\|[\theta]_i\,)\right],$$

where $\bar{\mu}_1(\cdot) = [\mu]_1(\cdot)$ and $\bar{\mu}_2(\cdot) = [\mu]_{1,2}(\cdot\,\big|\bar{X}_1)$. Note that $\bar{\mu}_2$ is a *random* measure, and that the integration with respect to $[\mu]_1$ in the chain rule is accounted for by the expectation operator E.

There is an obvious extension to any finite collection of independent random variables, which we state now as a lemma.

Lemma 7 *Let $\{X_i, i \in \mathbb{N}\}$ be iid S-valued random variables with distribution θ. Let $n \in \mathbb{N}$. If $f : S^n \to \mathbb{R}$ is bounded and measurable, then*

$$-\log Ee^{-f(X_1,\ldots,X_n)} = \inf E\left[f(\bar{X}_1, \ldots, \bar{X}_n) + \sum_{i=1}^{n} R\,(\bar{\mu}_i^n\,\|\theta\,)\right],$$

where the infimum is over all collections of random probability measures $\{\bar{\mu}_i^n, i \in \{1, \ldots, n\}\}$ that satisfy

1. *$\bar{\mu}_i^n$ is measurable with respect to the σ-algebra generated by $\bar{X}_1, \ldots, \bar{X}_{i-1}$, and*

2. the conditional distribution of $\bar{X}_i$, given $\bar{X}_1, \ldots, \bar{X}_{i-1}$, is $\bar{\mu}_i^n$.

We consider $\left\{\bar{X}_j, j = 1, \ldots, n\right\}$ to be a *controlled* version of the original sequence $\{X_j, j = 1, \ldots, n\}$, with the control $\bar{\mu}_j^n$ selecting the (conditional) distribution of $\bar{X}_j$. With the appropriate large deviation scaling, the representation becomes

$$-\frac{1}{n}\log Ee^{-nf(X_1,\ldots,X_n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[f(\bar{X}_1,\ldots,\bar{X}_n) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n \,\|\theta\right)\right].$$

Notational convention. Throughout these notes, we will use overbars to indicate the controlled analogue of any uncontrolled process.

The chain rule can be used in a similar fashion to immediately construct convenient representations for virtually any discrete time problem, and we do so for various classes of noise models later in the notes. In general, each of the "driving noises" needed to construct the system is replaced by a controlled analogue, where the control can in principle depend on all other previously defined controlled noises, and a relative entropy cost is paid for the different between the (perhaps conditional) distribution used to create the controlled noise and the distribution used to construct the original noise.

3 An overview of the weak convergence approach

Before getting into particular examples, we pause to comment on the main ideas that appear in every application of the representations to prove large deviation properties. With an appropriate scaling parameter (e.g., n), the prelimit Laplace quantity takes the form

$$-\frac{1}{n}\log\int_{S_n} e^{-nF(G^n(x))}\theta^n(dx),$$

and we need to prove convergence to

$$\inf_{y\in S}\left[F(y) + I(y)\right],$$

where $G^n : S_n \to S$ (here S_n is a Polish space indexed by n). Thus we prove

$$\inf_{\mu\in\mathcal{P}(S_n)}\left[\int_{S_n} F(G^n(x))\mu(dx) + R\left(\mu\,\|\theta^n\right)\right] \to \inf_{y\in S}\left[F(y) + I(y)\right].$$

The "high level" relative entropy $R\left(\mu\,\|\theta^n\right)$ should first be rewritten as dictated by the structure of the problem. Suppose for an optimizing sequence μ^n that $G^n(x)$ (under μ^n) converges in some sense to a point y. Then the total relative entropy cost $R\left(\mu^n\,\|\theta^n\right)$ must converge to $I(y)$, and this will identify the rate function.

In practice the proof is split into upper and lower bounds that are not symmetric, and analogous to what is sometimes called Γ-convergence. The lower bound

$$\liminf_{n\to\infty} \inf_{\mu\in\mathcal{P}(S_n)} \left[\int_{S_n} F(G^n(x))\mu(dx) + R\left(\mu \,\|\theta^n\right) \right] \geq \inf_{y\in S} \left[F(y) + I(y)\right]$$

gets to assume that $R\left(\mu \,\|\theta^n\right)$ is uniformly bounded (since otherwise boundedness of F would imply the left side tends to infinity). This bound is then used to prove some kind of tightness of the controls and controlled processes (the definition of tightness is recalled below), and the key issue will be to relate the weak limits of the controls and controlled processes. Because of the direction of the inequality and convexity properties (e.g., convexity of relative entropy), things like lower semicontinuity, Jensen's inequality and Fatou's Lemma are useful.

The argument for the upper bound

$$\limsup_{n\to\infty} \inf_{\mu\in\mathcal{P}(S_n)} \left[\int_{S_n} F(G^n(x))\mu(dx) + R\left(\mu \,\|\theta^n\right) \right] \leq \inf_{y\in S} \left[F(y) + I(y)\right]$$

generally starts with a near minimizer for the right hand side. One must show how to adapt this near minimizer to design a control for the prelimit controlled processes for which one can show convergence of controlled processes and costs, and in particular the convergence of the relative entropy cost to the rate function. While this might appear easier in that one is not dealing with a sequence of more-or-less arbitrary controls subject to a relative entropy bound, one also does not have Fatou, Jensen, etc., to help out. Thus it can happen that more involved constructions are needed to make the convergence go through.

4 Tightness and tightness functions

We end this lecture by stating some results that will be helpful in establishing tightness of controls and controlled processes. Let A be an index set and let $\{\lambda_a, a \in A\} \subset \mathcal{P}(S)$, where S is a Polish space. Recall that the collection $\{\lambda_a, a \in A\}$ is said to be *tight* if for all $\varepsilon > 0$ there is compact $K_\varepsilon \subset S$ such that $\inf\{\lambda_a(K_\varepsilon) : a \in A\} \geq 1 - \varepsilon$. If random variables $\{X_a, a \in A\}$ have the distributions $\{\lambda_a, a \in A\}$, we say $\{X_a, a \in A\}$ is tight if and only if $\{\lambda_a, a \in A\}$ is tight. According to Prohorov's Theorem, $\{\lambda_a, a \in A\}$ is precompact in the topology of weak convergence if and only if it is tight.

The notion of a tightness function will be useful. A measurable function $g : S \to [0, \infty]$ is called a tightness function if it has precompact level sets: for every $M \in [0, \infty)$ the set $\{x \in S : g(x) \leq M\}$ has compact closure. Thus rate functions are tightness functions. We have the following elementary result.

Lemma 8 *A collection $\{\lambda_a, a \in A\} \subset \mathcal{P}(S)$ is tight if and only if there is a tightness function g such that $\sup_{a\in A} \int_S g(x)\lambda_a(dx) < \infty$.*

Proof. The "$\Leftarrow$" direction follows directly from Chebyshev's inequality. To argue the "$\Rightarrow$" direction, let K_ε satisfy the requirement in the definition of tightness for $\{\lambda_a, a \in A\}$, and let K_ε^c denote the complement. Then

$$g(x) = \sum_{i=1}^{\infty} 1_{K^c_{2^{-i}}}(x)$$

serves as a tightness function with the desired properties. □

The next result shows that tightness functions have a useful "bootstrap" property.

Lemma 9 *Let g be a tightness function on S. Define $G : \mathcal{P}(S) \to [0, \infty]$ by*

$$G(\mu) = \int_S g(x)\mu(dx).$$

Then

- *for each $M < \infty$ the set $\{\mu \in \mathcal{P}(S) : G(\mu) \leq M\}$ is tight (and hence precompact), and*
- *G is a tightness function on $\mathcal{P}(S)$.*

Proof. The claims follow directly from Prohorov's Theorem and the preceding lemma. □

Lemma 10 *Let $\{\Lambda_a, a \in A\}$ be random variables taking values in $\mathcal{P}(S)$ (i.e., random probability measures), and let $\lambda_a = E\Lambda_a$. Then $\{\Lambda_a, a \in A\}$ is tight if and only if $\{\lambda_a, a \in A\}$ is tight. In other words, a collection of random probability measures are tight (as random variables!) if their "means" are tight as deterministic probability measures.*

Proof. Let η_a denote the distribution of Λ_a on $\mathcal{P}(S)$ and let $\varepsilon > 0$ be given. Assuming that the random measures $\{\Lambda_a, a \in A\}$ are tight, there is a compact set $K \subset \mathcal{P}(S)$ such that $\eta_a(K^c) \leq \varepsilon$. Since K is compact, there is $K_1 \subset \mathcal{P}(S)$ such that $\lambda \in K$ implies $\lambda(K_1^c) \leq \varepsilon$. Therefore

$$\begin{aligned}
\lambda_a(K_1^c) &= \int_{\mathcal{P}(S)} \lambda(K_1^c)\eta_a(d\lambda) \\
&= \int_K \lambda(K_1^c)\eta_a(d\lambda) + \int_{K^c} \lambda(K_1^c)\eta_a(d\lambda) \\
&\leq 2\varepsilon.
\end{aligned}$$

Thus $\{\lambda_a, a \in A\}$ is tight.

To prove the "$\Leftarrow$" direction, it suffices to find a tightness function $\bar{G} : \mathcal{P}(S) \to [0, \infty]$ such that

$$\sup_{a \in A} E\bar{G}(\Lambda_a) < \infty.$$

Since $\{\lambda_a, a \in A\}$ is tight we know there is a tightness function $\bar{g} : S \to [0, \infty]$ such that

$$\sup_{a \in A} \int_S \bar{g}(x)\lambda_a(dx) < \infty.$$

Letting $\bar{G}(\rho) = \int_S \bar{g}(x)\rho(dx)$, we observe that by construction $E\int_S \bar{g}(x)\Lambda_a(dx) = \int_S \bar{g}(x)\lambda_a(dx)$, and therefore

$$\sup_{a \in A} E\bar{G}(\Lambda_a) = \sup_{a \in A} E\int_S \bar{g}(x)\Lambda_a(dx) = \sup_{a \in A} \int_S \bar{g}(x)\lambda_a(dx) < \infty.$$

By the previous lemma $\bar{G}$ is a tightness function on $\mathcal{P}(S)$, which completes the proof. $\square$

Lecture 3: Canonical Problem I – Sanov's Theorem

In this lecture we prove Sanov's Theorem, one of the basic results of the theory, and show how under additional assumptions Cramér's Theorem follows. This might seem strange, since in many ways Cramér's Theorem appears simpler. The proofs for more complicated models that appear later in the notes will use many of the same arguments.

1 Sanov's Theorem

First we recall the statement of the Glivenko-Cantelli Lemma.

Lemma 1 (Glivenko-Cantelli) *Let $\{X_i, i \in \mathbb{N}\}$ be iid S-valued random variables with distribution γ, and let L^n be the empirical measure of the first n variables:*

$$L^n(dx) \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(dx).$$

Then w.p.1, L^n converges in the weak topology to γ.

The proof is a special case of the arguments we will use for Sanov's Theorem, and in particular follows from Lemmas 4 and 5. Sanov's Theorem itself is the large deviation refinement of this LLN result.

Lemma 2 *For each $\gamma \in \mathcal{P}(S)$ $R(\cdot\|\gamma)$ has compact level sets.*

Proof. Let $\{\mu_n, n \in \mathbb{N}\}$ be any sequence in $\mathcal{P}(S)$ satisfying $\sup_{n\in\mathbb{N}} R(\mu_n\|\gamma) \le M < \infty$. It follows from the variational formula for exponential integrals (see Lecture 2) that for bounded measurable ψ mapping S into $\mathbb{R}$ and $n \in \mathbb{N}$,

$$\int_S \psi \, d\mu_n - \log \int_S e^{\psi} \, d\gamma \le R(\mu_n\|\gamma) \le M.$$

Let $\delta > 0$ and $\varepsilon > 0$ be given. The tightness of γ guarantees that there exists a compact set K such that $\gamma(K^c) \le \varepsilon$. Substituting into the last display the function ψ that equals 0 on K and equals $\log(1 + 1/\varepsilon)$ on K^c, we have for each $n \in \mathbb{N}$

$$\begin{aligned}
\mu_n(K^c) &\le \frac{1}{\log(1+1/\varepsilon)} \left(M + \log \left[\gamma(K) + \left(1 + \frac{1}{\varepsilon}\right) \gamma(K^c) \right] \right) \\
&= \frac{1}{\log(1+1/\varepsilon)} \left(M + \log \left[1 + \frac{1}{\varepsilon} \gamma(K^c) \right] \right) \\
&\le \frac{1}{\log(1+1/\varepsilon)} (M + \log 2).
\end{aligned}$$

Since $\varepsilon > 0$ can be chosen so that $(M + \log 2)/\log(1 + 1/\varepsilon) \le \delta$, this formula implies that $\{\mu_n\}$ is tight. By Prohorov's Theorem there exists $\mu \in \mathcal{P}(S)$

and a subsequence of $n \in \mathbb{N}$ such that $\mu_n \Longrightarrow \mu$. The lower semicontinuity of $R(\cdot\|\gamma)$ yields

$$R(\mu\|\gamma) \leq \liminf_{n\to\infty} R(\mu_n\|\gamma) \leq M.$$

This completes the proof that $\{\mu \in \mathcal{P}(S) : R(\mu\|\gamma) \leq M\}$ is compact. □

Theorem 3 (SANOV'S THEOREM) *Let $\{X_i, i \in \mathbb{N}\}$ be iid S-valued random variables with distribution γ. Then $\{L^n, n \in \mathbb{N}\}$ satisfies the LDP on $\mathcal{P}(S)$ with rate function $I(\mu) = R(\mu\,\|\gamma)$.*

Using the Laplace Principle formulation, it is enough to show that

$$\lim_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} = \inf_{\mu\in\mathcal{P}(S)} \left[F(\mu) + R(\mu\,\|\gamma)\right]$$

for any bounded and continuous function F on $\mathcal{P}(S)$. Using the representation

$$-\frac{1}{n}\log Ee^{-nf(X_1,\ldots,X_n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[f(\bar{X}_1,\ldots,\bar{X}_n) + \frac{1}{n}\sum_{i=1}^n R(\bar{\mu}_i^n\,\|\gamma)\right],$$

where $\bar{\mu}_i^n$ selects the distribution of $\bar{X}_i$, given $\bar{X}_j, j = 1,\ldots,i-1$. The choice $f(x_1,\ldots,x_n) = F\left(\sum_{i=1}^n \delta_{x_i}(dx)/n\right)$ gives

$$-\frac{1}{n}\log Ee^{-nF(L^n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R(\bar{\mu}_i^n\,\|\gamma)\right].$$

Thus we need to show that

$$\inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R(\bar{\mu}_i^n\,\|\gamma)\right] \to \inf_{\mu\in\mathcal{P}(S)} \left[F(\mu) + R(\mu\,\|\gamma)\right].$$

Since F is bounded, the infimum in the representation is always bounded above by $\|F\|_\infty < \infty$. **It follows that without loss we can always restrict to controls for which the relative entropy cost is bounded by $2\|F\|_\infty < \infty$.**

1.1 Tightness and weak convergence

The bound on relative entropy costs is all that is *available*, but also all that is *needed*, to prove tightness.

Lemma 4 *Let $\{\bar{\mu}_i^n\}$ be a collection of controls for which $E\left[\frac{1}{n}\sum_{i=1}^n R(\bar{\mu}_i^n\,\|\gamma)\right]$ is uniformly bounded, and let $\hat{\mu}^n = \frac{1}{n}\sum_{i=1}^n \bar{\mu}_i^n$. Then $\left\{(\bar{L}^n, \hat{\mu}^n), n \in \mathbb{N}\right\}$ is tight.*

Proof. It follows from the convexity of relative entropy and Jensen's inequality that $E\left[\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \geq E\left[R\left(\hat{\mu}^n \,\|\gamma\right)\right]$. Since $\mu \to R\left(\mu \,\|\gamma\right)$ has compact level sets it is a tightness function, and so both $\{\hat{\mu}^n, n \in \mathbb{N}\}$ and $\{E\hat{\mu}^n, n \in \mathbb{N}\}$ are tight. Since $\bar{\mu}_i^n$ is the conditional distribution used to select $\bar{X}_i$, for any measurable function

$$E\int_S f(x)\bar{L}^n(dx) = E\frac{1}{n}\sum_{i=1}^n f(\bar{X}_i) = E\frac{1}{n}\sum_{i=1}^n \int_S f(x)\bar{\mu}_i^n(dx) = E\int_S f(x)\hat{\mu}^n(dx).$$

Thus $E\bar{L}^n = E\hat{\mu}^n$, and so $\left\{\bar{L}^n, n \in \mathbb{N}\right\}$ and hence $\left\{(\bar{L}^n, \hat{\mu}^n), n \in \mathbb{N}\right\}$ are tight. □

Thus $(\bar{L}^n, \hat{\mu}^n)$ will converge in distribution, at least along subsequences. To prove the LDP we need to relate the limits of the controls $\hat{\mu}^n$ and the controlled process $\bar{L}^n$.

Lemma 5 *Suppose* $\left\{(\bar{L}^n, \hat{\mu}^n), n \in \mathbb{N}\right\}$ *converges along a subsequence to* $(\bar{L}, \hat{\mu})$. *Then* $\bar{L} = \hat{\mu}$.

The proof of this result, which is a martingale version of the proof of the Glivenko-Cantelli Lemma, will be given after we complete Sanov's Theorem.

1.2 Lower bound

As remarked at the end of Lecture 2, the proof is partitioned into upper and lower bounds. Owing to an intervening minus sign, the large deviation upper bound corresponds a lower bound on the representation, and vice versa. For each $\varepsilon > 0$ let $\{\bar{\mu}_i^n, i \in \{1, \ldots, n\}\}$ and $\left\{\bar{X}_i, i \in \{1, \ldots, n\}\right\}$ satisfy

$$-\frac{1}{n}\log Ee^{-nF(L^n)} + \varepsilon \geq E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right].$$

Consider any subsequence of $\left\{(\bar{L}^n, \hat{\mu}^n), n \in \mathbb{N}\right\}$. Owing to tightness we can extract a further subsequence that converges weakly. If the lower bound is demonstrated for this subsequence, the standard argument by contradiction establishes the lower bound for the original sequence. To simplify notation, we denote the convergent subsequence by n, and its limit by $\left(\bar{L}, \hat{\mu}\right)$. According to Lemma 5, $\bar{L} = \hat{\mu}$ a.s. In the following display, we using Jensen's inequality and convexity of relative entropy for the second inequality, the convergence in distribution, lower semicontinuity of relative entropy and

Fatou's Lemma for the third inequality, and $\bar{L} = \hat{\mu}$ a.s. for the last:

$$\begin{aligned}
\liminf_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} + \varepsilon &\geq \liminf_{n\to\infty} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \\
&\geq \liminf_{n\to\infty} E\left[F\left(\bar{L}^n\right) + R\left(\hat{\mu}^n \,\|\gamma\right)\right] \\
&\geq E\left[F\left(\bar{L}\right) + R\left(\hat{\mu} \,\|\gamma\right)\right] \\
&\geq \inf_{\mu\in\mathcal{P}(S)} \left[F(\mu) + R\left(\mu \,\|\gamma\right)\right].
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the result follows. □

1.3 Upper bound

Next we prove the reverse inequality. For $\varepsilon > 0$ let μ^* satisfy

$$\left[F(\mu^*) + R\left(\mu^* \,\|\gamma\right)\right] \leq \inf_{\mu\in\mathcal{P}(S)} \left[F(\mu) + R\left(\mu \,\|\gamma\right)\right] + \varepsilon.$$

Then let $\bar{\mu}_i^n = \mu^*$ for all $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$, so the $\bar{X}_i$ are iid with distribution μ^*. By either Lemma 5 or the ordinary Glivenko-Cantelli Lemma, the weak limit of $\bar{L}^n$ equals μ^*. The fact that this particular choice of $\{\bar{\mu}_i^n\}$ is not necessarily infimizing gives the first inequality below, and the convergence in distribution version of the Dominated Convergence Theorem gives the equality:

$$\begin{aligned}
\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} &\leq \limsup_{n\to\infty} E\left[F\left(\bar{L}^n\right) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n \,\|\gamma\right)\right] \\
&= \left[F\left(\mu^*\right) + R\left(\mu^* \,\|\gamma\right)\right] \\
&\leq \inf_{\mu\in\mathcal{P}(S)} \left[F(\mu) + R\left(\mu \,\|\gamma\right)\right] + \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the upper bound follows. □

Proof of Lemma 5. Since S is Polish there exists countable separating class $\{f_m, m \in \mathbb{N}\}$ of bounded, continuous functions. Define $K_m = \sup_{x\in S} |f_m(x)|$ and $\Delta_{m,i}^n = f_m\left(\bar{X}_i^n\right) - \int_S f_m(x)\,\bar{\mu}_i^n(dx)$. For any $\varepsilon > 0$

$$\begin{aligned}
&P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\int_S f_m(x)\,\delta_{\bar{X}_i^n}(dx) - \frac{1}{n}\sum_{i=1}^{n}\int_S f_m(x)\,\bar{\mu}_i^n(dx)\right| > \varepsilon\right] \\
&\qquad \leq \frac{1}{\varepsilon^2} E\left[\frac{1}{n^2}\sum_{i,j=1}^{n} \Delta_{m,i}^n \Delta_{m,j}^n\right].
\end{aligned}$$

Let $\mathcal{F}_j^n = \sigma(\bar{X}_i^n, i = 1, \ldots, j)$. By a standard conditioning argument, the off-diagonal terms vanish: for $i > j$

$$E\left[\Delta_{m,i}^n \Delta_{m,j}^n\right] = E\left[E\left[\Delta_{m,i}^n \Delta_{m,j}^n \middle| \mathcal{F}_i^n\right]\right] = E\left[E\left[\Delta_{m,i}^n \middle| \mathcal{F}_i^n\right] \Delta_{m,j}^n\right] = 0.$$

Since $|\Delta_{m,i}^n| \leq 2K_m$,

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\int_S f_m(x)\,\delta_{\bar{X}_i^n}(dx) - \frac{1}{n}\sum_{i=1}^{n}\int_S f_m(x)\,\bar{\mu}_i^n(dx)\right| > \varepsilon\right] \leq \frac{4K_m^2}{n\varepsilon^2}.$$

Since $(\bar{L}^n, \hat{\mu}^n) \Rightarrow \left(\bar{L}, \hat{\mu}\right)$ and $\varepsilon > 0$ is arbitrary, by Fatou's lemma

$$P\left[\int_S f_m(x)\,\bar{L}(dx) = \int_S f_m(x)\,\hat{\mu}(dx)\right] = 1.$$

Now use that $\{f_m, m \in \mathbb{N}\}$ is countable and separating to conclude $\bar{L} = \hat{\mu}$ w.p.1. □

2 Cramér's Theorem

Cramér's Theorem states the LDP for the empirical mean of $\mathbb{R}^d$-valued iid random variables: $S_n \doteq \frac{1}{n}(X_1 + \cdots + X_n)$. Of course one can recover the empirical mean from the empirical measure via $S_n = \int_{\mathbb{R}^d} yL^n(dy)$. If the underling distribution γ has compact support then the mapping $\mu \to \int_{\mathbb{R}^d} y\mu(dy)$ is continuous on a subset of $\mathcal{P}(\mathbb{R}^d)$ that contains L^n w.p.1, and the rate function I for $\{S_n, n \in \mathbb{N}\}$ follows directly from the contraction principle. For $\beta \in \mathbb{R}^d$,

$$I(\beta) = \inf\left\{R(\mu\,\|\gamma) : \int_{\mathbb{R}^d} y\mu(dy) = \beta\right\}. \tag{2}$$

However, in general the mapping $\mu \to \int_{\mathbb{R}^d} y\mu(dy)$ is not continuous, and the contraction principle does not suffice. The problem is that the conditions of Sanov's Theorem are too weak to force continuity with high probability. They are sufficient to imply tightness of controls, but no more. Once the conditions are appropriately strengthened, the weak convergence arguments can be carried out just as before, with the only difference being in the qualitative properties of the convergence. For $\alpha \in \mathbb{R}^d$ let

$$H(\alpha) = \log \int_{\mathbb{R}^d} e^{\langle \alpha, y\rangle}\gamma(dy).$$

Theorem 6 *Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of iid $\mathbb{R}^d$-valued random variables with common distribution γ, and let $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Assume that $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$. Then $\{S_n, n \in \mathbb{N}\}$ satisfies the LDP with rate function I defined as in (2).*

Proof. To apply the weak convergence we should consider $F : \mathbb{R}^d \to \mathbb{R}$ that is bounded and continuous, and calculate the limits of

$$-\frac{1}{n}\log Ee^{-nF\left(\int_{\mathbb{R}^d} yL^n(dy)\right)}. \tag{3}$$

Thus it is a special case of the representation used in Sanov's Theorem that is relevant, which is

$$\inf_{\{\bar{\mu}_i^n\}} E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n \,\|\gamma\right)\right].$$

Suppose we prove that $\left\{\bar{L}^n, n \in \mathbb{N}\right\}$ is uniformly integrable, in the sense that whenever the relative entropy cost is uniformly bounded,

$$\lim_{M\to\infty} \limsup_{n\to\infty} E\left[\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy)\right] = 0.$$

Then it will follow that

$$E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}^n(dy)\right)\right] \to E\left[F\left(\int_{\mathbb{R}^d} y\bar{L}(dy)\right)\right],$$

and the limit of (3) can be calculated using exactly the same argument as that used to prove Sanov's Theorem. The integrability should come from the bound on relative entropy costs and the new assumption $H(\alpha) < \infty$.

For $b \geq 0$ let $\ell(b) = b\log b - b + 1$. A bound that is used frequently in large deviations is that for $a \geq 0$, $b \geq 0$ and $\sigma \geq 1$

$$ab \leq e^{\sigma a} + \frac{1}{\sigma}\left(b\log b - b + 1\right) = e^{\sigma a} + \frac{1}{\sigma}\ell(b).$$

This follows from the fact that

$$\sup_{a\in\mathbb{R}}\left\{ab - e^{\sigma a}\right\} = \frac{b}{\sigma}\left(\log\frac{b}{\sigma} - 1\right) \leq \frac{1}{\sigma}\ell(b).$$

Thus if $\theta \in \mathcal{P}(\mathbb{R}^d)$ satisfies $\theta \ll \gamma$, then

$$\begin{aligned}
\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}}\theta(dy) &= \int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}} \frac{d\theta}{d\gamma}(y)\gamma(dy) \\
&\leq \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\geq M\}}\gamma(dy) + \frac{1}{\sigma}\int_{\mathbb{R}^d} \ell\left(\frac{d\theta}{d\gamma}(y)\right)\gamma(dy) \\
&= \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\geq M\}}\gamma(dy) + \frac{1}{\sigma}R\left(\theta\,\|\gamma\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
E\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}} \bar{L}^n(dy) &= E\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\|\geq M\}} \hat{\mu}^n(dy) \\
&\leq \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\|\geq M\}}\gamma(dy) + \frac{1}{\sigma}ER\left(\hat{\mu}^n\,\|\gamma\right).
\end{aligned}$$

Using the uniform bounds on the relative entropy and the fact that $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$, taking limits in the order $M \to \infty$ and then $\sigma \to \infty$ gives the uniform integrability. □

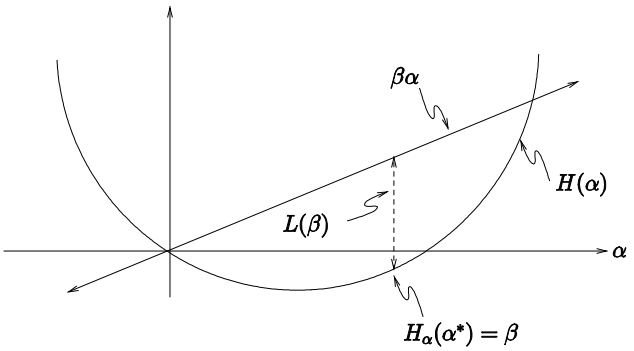

Figure 6: Geometry of the Legendre transform in one dimension

Remark 1 The "usual" form of the rate function in the statement of Cramér's Theorem is as the Legendre transform of H:

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H(\alpha) \right].$$

One can directly verify that the two coincide (see the end of Lecture 6). The representation in terms of relative entropy is often more useful. Also, we note that one can also prove the conclusion under the condition that for some $\delta > 0$, $H(\alpha) < \infty$ for all α with $\|\alpha\| \leq \delta$. However, this requires unbounded test functions F and a different line of argument for the lower bound (the large deviation upper bound).

Some examples of $H - L$ pairs are as follows.

Example 1 Suppose that X_1 is Bernoulli with

$$P\{X_1 = 0\} = 1 - p, \; P\{X_1 = 1\} = p$$

for $p \in (0, 1)$. Then

$$H(\alpha) = \log\left((1-p) + pe^{\alpha}\right)$$

and *(*with the understanding that $0 \log 0 = 0$*)*

$$L(\beta) = \begin{cases} \beta \log\left(\frac{\beta}{p}\right) + (1-\beta)\log\left(\frac{1-\beta}{1-p}\right) & \beta \in [0,1] \\ \infty & \beta \notin [0,1] \end{cases}.$$

Example 2 Suppose that X_1 is Poisson with parameter $\lambda > 0$, so that $P\{X_1 = n\} = e^{-\lambda}\lambda^n/n!$ for $n \in \mathbb{N}_0$. Then

$$H(\alpha) = \lambda\left(e^{\alpha} - 1\right)$$

and

$$L(\beta) = \begin{cases} \beta \log\left(\frac{\beta}{\lambda}\right) - \beta \; + \lambda & \beta \geq 0 \\ \infty & \beta < 0 \end{cases}.$$

Example 3 Suppose that X_1 is Gaussian $N(b, \sigma^2)$. Then

$$H(\alpha) = \frac{\alpha^2\sigma^2}{2} + \alpha b$$

and

$$L(\beta) = \frac{1}{2\sigma^2}(\beta - b)^2.$$

3 An application of the conditioning result

Recall that Theorem 2 of Lecture 2 asserts under some conditions that given a rare outcome, the minimizing point in the large deviation rate identifies the most likely way the rare event occurs. We will show in Lecture 6 that the minimizing measure in

$$\inf\left\{R(\mu\,\|\gamma) : \int_{\mathbb{R}^d} y\mu(dy) = \beta\right\}$$

is of the form $e^{\langle\alpha,x\rangle - H(\alpha)}\gamma(dx)$, where α is chosen so that

$$\int_{\mathbb{R}^d} ye^{\langle\alpha,x\rangle - H(\alpha)}\gamma(dx) = \beta.$$

Suppose that the X_i's in Sanov's Theorem are Poisson with $\lambda = 1$, and that we observe $S_n = b > 1$. Then by this conditioning result, given this observation we find with probability approaching 1 as $n \to \infty$ that L^n appears to be nearly Poisson with parameter b. Thus the observed mean in some sense gives information on the likely form of the entire empirical distribution.

Lecture 4: Canonical Problem II – Small Noise Diffusions

1 A Representation for Functionals of Brownian Motion

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{\mathcal{F}_t, 0 \leq t \leq 1\}$ be a right-continuous P-complete filtration on $(\Omega, \mathcal{F}, P)$. Suppose that $\{W(t), 0 \leq t \leq 1\}$ is a d-dimensional $\mathcal{F}_t$-Brownian motion, i.e., $W(t)$ is $\mathcal{F}_t$-measurable for every $t \in [0,1]$, and $W(t) - W(s)$ is independent of $\mathcal{F}_s$ for all $0 \leq s \leq t \leq 1$. A standard choice of $\mathcal{F}_t$ is the sigma-field $\sigma\{W(s) : 0 \leq s \leq t\}$, augmented with all P-null sets, i.e.,

$$\mathcal{F}_t^W \doteq \sigma\left\{\sigma\{W(s) : 0 \leq s \leq t\} \vee \mathcal{N}\right\},$$

where $\mathcal{N} = \{A \subset \Omega : \text{there is } B \in \mathcal{F} \text{ with } A \subset B \text{ and } P(B) = 0\}$.

Definition 3 An $\mathbb{R}^k$-valued stochastic process $\{v(t), 0 \leq t \leq 1\}$ on $(\Omega, \mathcal{F}, P)$ is said to be $\mathcal{F}_t$-progressively measurable if for every $t \in [0,1]$ the map $(s, \omega) \mapsto v(s, \omega)$ from $([0,t] \times \Omega, \mathcal{B}([0,t]) \otimes \mathcal{F}_t)$ to $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ is measurable. Let $\mathcal{U}$ denote the collection of all $\mathcal{F}_t$-progressively measurable processes $\{v(t), 0 \leq t \leq 1\}$ which satisfy $E(\int_0^1 ||v(t)||^2 dt) < \infty$.

The following representation theorem for bounded measurable functionals of a Brownian motion is analogous to the one stated in Lecture 3 for functionals of an iid sequence. In the representation, the controlled measures have been replaced by just a control process, and the relative entropy cost is the expected L^2 norm of this process. The representation is given for the interval $[0,1]$, but extends in a trivial way to any bounded interval. Let $\mathcal{C}([0,1] : \mathbb{R}^k)$ denote the space of $\mathbb{R}^k$-valued continuous functions on $[0,1]$. This space is equipped with the uniform metric, which makes it a Polish space.

Theorem 1 *Let f be a bounded Borel measurable function from $\mathcal{C}([0,1] : \mathbb{R}^k)$ to $\mathbb{R}$. Then*

$$-\log E e^{-f(W)} = \inf_{v \in \mathcal{U}} E\left[f\left(W + \int_0^{\cdot} v(s)ds\right) + \frac{1}{2}\int_0^1 ||v(s)||^2 ds\right]. \qquad (4)$$

The proof of this representation appears in [3]. The form of the representation closely parallels the corresponding discrete time result for product measure, reflecting the fact that Brownian motion is the integral of "white" noise, and progressive measurability is analogous to the fact that in the representation for iid noises $\bar{\mu}_i^n$ is allowed to depend on all controlled noises up to time $i-1$. In fact, if one were to replace W by the correponding

piecewise linear interpolation with interpolation interval $\delta > 0$ (which is equivalent to a collection of $1/\delta$ iid $N(0,\delta)$ random variables), and assume that the minimizing measures are Gaussian with means $\bar{v}_i^n$, the L^2 cost in (4) corresponds to $R\left(N(\bar{v}_i^n,\delta)\,\|N(0,\delta)\right) = \delta||\bar{v}_i^n||^2/2$. The assumption that one can restrict the discrete time measures to ones of the form $N(\bar{v}_i^n,\delta)$ is valid in the limit $\delta \to 0$, which is why the continuous time representation is in some ways simpler than thhe corresponding discrete time representation.

An analgous representation is proved in [43] and applied to problem of risk sensitive control, though the expectation E in that representation depends on v. Extensions to infinite dimensional Gaussian processes (e.g., cylindrical Brownian motion, Brownian sheet) and continuous time jump noises (Poisson random measures) appear in [5, 7, 9].

2 Large Deviation Theory of Small Noise Diffusions

The representation (4) is very convenient for a weak convergence large deviation analysis, and in many ways it makes the continuous time setting simpler than the corresponding discrete time setting. As an illustration of its use we prove a large deviation principle for a class of small noise diffusions. While fairly general, the assumptions on the coefficients are chosen to make the presentation simple, and can be significantly relaxed. We assume there is $C \in (0,\infty)$ such that $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ satisfy

$$||b(x) - b(y)|| + ||\sigma(x) - \sigma(y)|| \le C||x-y|| \quad \text{and} \quad ||b(x)|| + ||\sigma(x)|| \le C$$

for all $x, y \in \mathbb{R}^d$. Fix $x \in \mathbb{R}^d$, and for $\varepsilon > 0$ let $X^\varepsilon = \{X^\varepsilon(t), 0 \le t \le 1\}$ be the pathwise solution of the stochastic differential equation (SDE)

$$dX^\varepsilon(t) = b(X^\varepsilon(t))dt + \sqrt{\varepsilon}\sigma(X^\varepsilon(t))dW(t),\ X^\varepsilon(0) = x. \tag{5}$$

Let $\mathcal{A}_x([0,1] : \mathbb{R}^d)$ denote the space of $\mathbb{R}^d$-valued, absolutely continuous functions φ on $[0,1]$ with $\varphi(0) = x$. Also, for $\varphi \in \mathcal{A}_x([0,1] : \mathbb{R}^d)$, let

$$U_\varphi = \left\{ u \in L^2([0,1] : \mathbb{R}^d) : \varphi(t) = x + \int_0^t b(\varphi(s))ds + \int_0^t \sigma(\varphi(s))u(s)ds,\ t \in [0,1] \right\},$$

where $L^2([0,1] : \mathbb{R}^d)$ is the space of $\mathbb{R}^d$-valued square integrable functions on $[0,1]$. For all other $\varphi \in \mathcal{C}([0,1] : \mathbb{R}^d)$ let U_φ be the empty set. The following large deviation principle for such small noise diffusions is one of the classical results in the theory [34]. The infimum over the empty set is taken to be ∞.

Theorem 2 *The collection* $\{X^\varepsilon, \varepsilon \in (0,1)\}$ *satisfies the LDP on* $\mathcal{C}([0,1] : \mathbb{R}^d)$ *with rate function*

$$I(\varphi) \doteq \inf_{u\in U_\varphi} \left\{ \frac{1}{2}\int_0^1 ||u(t)||^2 dt \right\}.$$

To prove this theorem, we must show that for bounded and continuous $F : \mathcal{C}([0,1] : \mathbb{R}^d) \to \mathbb{R}$

$$\lim_{\varepsilon \to 0} -\varepsilon \log E e^{-\frac{1}{\varepsilon} F(X^\varepsilon)} = \inf_{\varphi \in \mathcal{C}([0,1]:\mathbb{R}^d)} \{F(\varphi) + I(\varphi)\}.$$

The first step will be to interpret $F(X^\varepsilon)$ as a bounded measurable function of W. From the unique pathwise solvability of the SDE (5), it follows that for each $\varepsilon > 0$ there is a measurable map $\mathcal{G}^\varepsilon : \mathcal{C}([0,1] : \mathbb{R}^k) \to \mathcal{C}([0,1] : \mathbb{R}^d)$ such that $X^\varepsilon = \mathcal{G}^\varepsilon(\sqrt{\varepsilon} W)$. Moreover, if $v \in \mathcal{U}$, then Girsanov's Theorem and the pathwise uniqueness implies $\mathcal{G}^\varepsilon(\sqrt{\varepsilon} W(\cdot) + \int_0^\cdot v(s)ds)$ is the unique solution to

$$d\bar{X}^\varepsilon(t) = b(\bar{X}^\varepsilon(t))dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon(t))dW(t) + \sigma(\bar{X}^\varepsilon(t))v(t)dt, \ \bar{X}^\varepsilon(0) = x. \quad (6)$$

This implies the control representation

$$\begin{aligned}
&-\varepsilon \log E e^{-\frac{1}{\varepsilon} F(X^\varepsilon)} \\
&= -\varepsilon \log e^{-\frac{1}{\varepsilon} F \circ \mathcal{G}^\varepsilon(\sqrt{\varepsilon} W)} \\
&= \inf_{v \in \mathcal{U}} E \left\{ F \circ \mathcal{G}^\varepsilon \left(\sqrt{\varepsilon} W(\cdot) + \int_0^\cdot v(s)ds \right) + \frac{1}{2} \int_0^1 ||v(s)||^2 ds \right\} \\
&= \inf_{v \in \mathcal{U}} E \left\{ F\left(\bar{X}^\varepsilon\right) + \frac{1}{2} \int_0^1 ||v(s)||^2 ds \right\}.
\end{aligned}$$

We would like to consider v as taking values in $L^2([0,1] : \mathbb{R}^d)$ with the weak topology. However, this space is not metrizable as a Polish space. Since F is bounded, given any $\delta > 0$ we can find $M_\delta < \infty$ and a control v that comes within δ of the infimum and such that $\frac{1}{2} \int_0^1 ||v(s)||^2 ds \leq M_\delta$ w.p.1. Thus we can assume that the controls take values in the compact Polish space S_{M_δ} where

$$S_M = \left\{ \phi \in L^2([0,1] : \mathbb{R}^d) : \frac{1}{2} \int_0^1 ||\phi(s)||^2 ds \leq M \right\}$$

and the weak topology on $L^2([0,1] : \mathbb{R}^d)$ is used. The proof, which we omit, starts with a nearly minimizing control, for which the relative entropy cost is necessarily bounded. One uses Chebyshev's inequality to show that the set where $\int_0^1 ||v(s)||^2 ds$ is large is small, and then uses a stopping time to modify the definition of v on this set. Since F is bounded the resulting total cost is still close to the infimum.

We will follow the same scheme of proof as in Sanov's Theorem. Thus we first prove a tightness result and show how to relate the weak limits of controls and controlled processes. The proof of the lower bound (which corresponds to the large deviation upper bound) as well as the proof that I is a rate function follows, and we conclude with the proof of the upper bound (the large deviation lower bound).

2.1 Tightness and weak convergence

Lemma 3 *Consider any collection of controls* $\{v^\varepsilon\} \subset \mathcal{U}$ *for which* $\frac{1}{2}\int_0^1 ||v^\varepsilon||^2 ds$ *is uniformly bounded by* $M < \infty$*, and define* $\bar{X}^\varepsilon$ *by (6) with* $v = v^\varepsilon$*. Then* $\left\{(\bar{X}^\varepsilon, v^\varepsilon), \varepsilon \in (0,1)\right\}$ *is tight.*

Proof. Tightness of $\{v^\varepsilon\}$ is automatic since S_M is compact. For the tightness of $\left\{\bar{X}^\varepsilon\right\}$, note that

$$\bar{X}^\varepsilon(t) - x = \int_0^t b(\bar{X}^\varepsilon(s))ds + \sqrt{\varepsilon}\int_0^t \sigma(\bar{X}^\varepsilon(s))dW(s) + \int_0^t \sigma(\bar{X}^\varepsilon(s))v^\varepsilon(s)ds.$$

The first and second terms are tight since b and σ are bounded using, e.g., the Kolmogorov-Čentsov criteria. Tightness of the third follows from the boundedness of σ, and since for $0 \le s \le t \le 1$

$$\int_s^t ||v^\varepsilon(r)||dr \le (t-s)^{1/2}\left(\int_0^1 ||v^\varepsilon(r)||^2 dr\right)^{1/2} \le (t-s)^{1/2}M^{1/2}.$$

□

Lemma 4 *Suppose for each* $\varepsilon \in (0,1)$ *that* $(\bar{X}^\varepsilon, v^\varepsilon)$ *solves (6), and that* $(\bar{X}^\varepsilon, v^\varepsilon)$ *converges weakly to* $(\bar{X}, v)$*. Then w.p.1*

$$\bar{X}(t) - x = \int_0^t b(\bar{X}(s))ds + \int_0^t \sigma(\bar{X}(s))v(s)ds. \tag{7}$$

Proof. By a standard martingale bound the stochastic integral converges to 0 as $\varepsilon \to 0$. The only remaining issue is to check that $\int_0^t \sigma(\bar{X}^\varepsilon(s))v^\varepsilon(s)ds$ converges to $\int_0^t \sigma(\bar{X}(s))v(s)ds$. By the Skorokod representation we can assume the convergence is w.p.1, where the topology of uniform convergence is used for $\bar{X}^\varepsilon$ and the weak topology on L^2 is used for v^ε. We have

$$\begin{aligned}&\int_0^t \sigma(\bar{X}^\varepsilon(s))v^\varepsilon(s)ds - \int_0^t \sigma(\bar{X}(s))v(s)ds \\ &= \int_0^t \left[\sigma(\bar{X}^\varepsilon(s)) - \sigma(\bar{X}(s))\right]v^\varepsilon(s)ds + \int_0^t \sigma(\bar{X}(s))\left[v^\varepsilon(s) - v(s)\right]ds.\end{aligned}$$

The first term tends to zero by Hölder's inequality and the second to zero since $\sigma(\bar{X}(\cdot)) \in L^2$. □

2.2 Lower bound

For each $\delta > 0$, let $\left\{(\bar{X}^\varepsilon, v^\varepsilon), \varepsilon \in (0,1)\right\}$ satisfy

$$-\varepsilon \log E e^{-\frac{1}{\varepsilon}F(X^\varepsilon)} + \delta \ge E\left[F\left(\bar{X}^\varepsilon\right) + \frac{1}{2}\int_0^1 ||v^\varepsilon(s)||^2 ds\right]$$

and $v^\varepsilon \in S_{M_\delta}$ w.p.1. Consider a weakly converging subsequence, with limit $(\bar{X}, v)$. Then using the definition of I and (7)

$$\begin{aligned}\liminf_{\varepsilon\to 0} -\varepsilon \log E e^{-\frac{1}{\varepsilon}F(X^\varepsilon)} + \delta &\geq \liminf_{\varepsilon\to 0} E\left[F\left(\bar{X}^\varepsilon\right) + \frac{1}{2}\int_0^1 ||v^\varepsilon(s)||^2 ds\right] \\ &\geq E\left[F\left(\bar{X}\right) + \frac{1}{2}\int_0^1 ||v(s)||^2 ds\right] \\ &\geq \inf_{\varphi\in\mathcal{C}([0,1]:\mathbb{R})} \left\{F(\varphi) + I(\varphi)\right\}.\end{aligned}$$

The usual argument by contradiction establishes the bound for the full sequence $\varepsilon \in (0,1)$. Now let $\delta \downarrow 0$. □

2.3 Upper bound

For $\delta > 0$ choose $\varphi^* \in \mathcal{C}([0,1] : \mathbb{R}^d)$ such that

$$F(\varphi^*) + I(\varphi^*) \leq \inf_{\varphi\in\mathcal{C}([0,1]:\mathbb{R}^d)} \left\{F(\varphi) + I(\varphi)\right\} + \delta.$$

Let $u \in \mathcal{S}_{\varphi^*}$ be such that $\frac{1}{2}\int_0^1 ||u(s)||^2 ds \leq I(\varphi^*) + \delta$ and let $\bar{X}^\varepsilon$ be the unique solution of (6) when we replace v on the right side of the equation by u. By the results on tightness and weak convergence (Lemmas 3 and 4) $\bar{X}^\varepsilon$ converges in probability to φ^*. Thus

$$\begin{aligned}\limsup_{\varepsilon\to 0} -\varepsilon \log E e^{-\frac{1}{\varepsilon}F(X^\varepsilon)} &= \limsup_{\varepsilon\to 0} \inf_{v\in\mathcal{U}} E\left[F\left(\bar{X}^\varepsilon\right) + \frac{1}{2}\int_0^1 ||v(s)||^2 ds\right] \\ &\leq \limsup_{\varepsilon\to 0} E\left[F\left(\bar{X}^\varepsilon\right) + \frac{1}{2}\int_0^1 ||u(s)||^2 ds\right] \\ &= F\left(\varphi^*\right) + \frac{1}{2}\int_0^1 ||u(s)||^2 ds \\ &\leq F\left(\varphi^*\right) + I\left(\varphi^*\right) + \delta \\ &\leq \inf_{\varphi\in\mathcal{C}([0,1]:\mathbb{R})} \left\{F\left(\varphi\right) + I\left(\varphi\right)\right\} + 2\delta.\end{aligned}$$

Since $\delta > 0$ is arbitrary, the upper bound follows. □

Note that when $\sigma(x)\sigma(x)^T > 0$ for all $x \in \mathbb{R}^d$, we can express I in the calculus of variations form

$$I(\varphi) = \int_0^1 L(\varphi, \dot{\varphi}) ds, \quad L(x,\beta) = \frac{1}{2}\left\langle (\beta - b(x)), [\sigma(x)\sigma(x)^T]^{-1}(\beta - b(x))\right\rangle.$$

Lecture 5: Freidlin-Wentsell Theory and Moderate Deviations

This lecture discusses two independent topics–the Freidlin-Wentsell theory and moderate deviations. The Freidlin-Wentsell theory leverages finite time large deviation estimates into a set of results on the metastability properties of small noise processes. We will not have time to go into this topic in detail, but it is one of the main motivations for the finite time estimates that are discussed in detail. Moderate deviation theory fills the gap between diffusion approximations and large deviation theory, and has some potential applications in Monte Carlo.

1 Freidlin-Wentsell theory for problems on unbounded time intervals

In Examples 1 and 4 of Lecture 1 there were "small noise" process level large deviation problems, where the event of interest involved a potentially unbounded time interval. We commented that for those problems one could essentially reduce to the case of a finite time analysis, since the probability that the event happened and took longer than T could be shown (using only finite time interval estimates) to have decay rate $K(T)$, where $K(T) \to \infty$ as $T \to \infty$. There are, however, very important problems also of interest on an a priori unbounded time interval which do not allow for such a simple reduction to finite time estimates. The key difference between the two types of problems is the presence of a rest point for the LLN limit dynamics (the paths which make the rate function zero) within the domain of interest.

One of the main achievements of large deviation theory is a collection of results, commonly referred to as the "Freidlin-Wentsell theory," which allow one to knit together finite time estimates with stopping time arguments to analyze these problems. The methods apply quite broadly, once one has the appropriate finite time estimates. They apply to all the process level large deviation models we consider, and even have extensions to non-Markovian models if the estimates are suitably uniform in the conditioning data [18]. To simplify we will consider only Markov process, and make assumptions on the rate function similar to those in the original work [34]. However, these assumptions can be relaxed considerably, as can virtually all the conditions we will impose. In particular, the assumption that the function L is finite is restrictive and can be relaxed as in [18], where one assumes a "controllable with small cost" property near the stable point.

Thus we assume that for each $x_0 \in \mathbb{R}^d$ and each $T \in (0, \infty)$ the family of Markov process $\{X^n, n \in \mathbb{N}\}$, conditioned on $X^n(0) = x_0$, satisfies the LDP

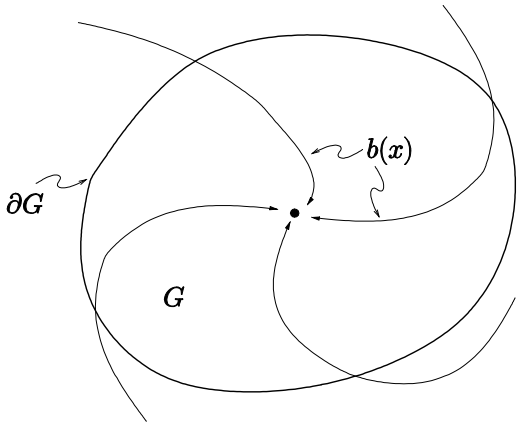

Figure 7: The domain G and the flow under b.

with a rate function of the form

$$I_{T,x_0}(\phi) = \begin{cases} \int_0^T L(\phi(s), \dot{\phi}(s))ds & \text{if } \phi \in \mathcal{A}_{x_0}([0,T] : \mathbb{R}^d), \\ \infty & \text{otherwise.} \end{cases}$$

We also assume that $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is continuous, convex in β for each fixed x, and uniformly superlinear:

$$\lim_{c \to \infty} \inf_{x \in \mathbb{R}^d} \inf_{\beta : ||\beta|| = c} \frac{1}{c} L(x, \beta) = \infty.$$

This latter condition guarantees that I_{T,x_0} has compact level sets. We will want the large deviation estimates to be uniform in the initial condition in the sense of [34, Section 3.3]. This uniformity condition generally holds, at least in finite dimensional settings. Finally we assume that for each $x \in \mathbb{R}^d$ there is a unique $b(x) \in \mathbb{R}^d$ such that $L(x, b(x)) = 0$, and that b is Lipschitz continuous. Thus the solutions to $\dot{\eta} = b(\eta)$ are the LLN limits for X^n as $n \to \infty$ when $X^n(0) = \eta(0) = x_0$.

We take as given a bounded, open, and connected set $G \subset \mathbb{R}^d$ with smooth boundary, and assume that there is a unique attractor for $\dot{\eta} = b(\eta)$ in G, which for convenience we label as 0. To be precise, it is assumed that if $n(x)$ is the outward normal to G at $x \in \partial G$, then $\sup_{x \in \partial G} \langle b(x), n(x) \rangle < 0$, and all solutions to $\dot{\eta} = b(\eta)$ starting at a point inside G converge to 0 as $t \to \infty$. See Figure 1.

The quasipotential with respect to 0 is defined in this context by

$$Q(x) = \inf \{ I_{T,0}(\phi) : \phi(T) = x, T < \infty \}.$$

Owing to the fact that 0 is stable, one can show that $\inf \{ I_{T,0}(\phi) : \phi(T) = x \}$ is decreasing to some strictly positive value as $T \to \infty$. This function gives

the approximate rate of decay for the probability to reach a neighborhood of x, after starting at the stable point 0 and over a long but finite time interval $[0, T]$. Under the given assumptions, it is easy to show that Q is continuous in G and strictly positive when $x \neq 0$.

Example 1 Suppose $b(x) = -DU(x) + \ell(x)$ where $\langle DU(x), \ell(x) \rangle = 0$, and that $L(x, \beta) = (\beta - b(x))^2/2$ (as would be true, for example, in the diffusion case). Then under the conditions on b stated above U will have a local minimum at $x = 0$, which is the only local minimum in G, and a standard verification argument will be used in Lecture 7 to show that $Q(x) = 2U(x)$. There are other examples for which the quasipotential takes an explicit form, including non-Gaussian noise and even infinite dimensional examples.

1.1 The exit location and the mean exit time

Among the many results describing the behavior of X^n over large time intervals we mention two, the first concerned with where the process will leave G, and the second with how long that will take. We first state the results and then suggest why they are valid. The results correspond to [34, Theorems 4.2.1 and 4.4.1.].

Theorem 1 *Suppose that there is a unique point $y \in \partial G$ such that $Q(y) = \inf \{Q(z) : z \in \partial G\}$. Let $\tau^n = \inf \{t : X^n(t) \in \partial G\}$. Then for all $\delta > 0$ and $x_0 \in G$*

$$\lim_{n \to \infty} P_{x_0} \{\|X^n(\tau^n) - y\| > \delta\} = 0.$$

Theorem 2 *Let $\tau^n = \inf \{t : X^n(t) \in \partial G\}$. Then for all $x_0 \in G$*

$$\lim_{n \to \infty} \frac{1}{n} \log E_{x_0} \tau^n = \inf \{Q(z) : z \in \partial G\}.$$

The proof of both these results use stopping times and the strong Markov property. However, the main idea is that a small neighborhood of the rest point should act like a renewal point.

Although the process may start at $x_0 \neq 0$, which probability tending to one it will approach the neighborhood of the rest point before exiting. Once near the rest point, occasional bursts of noise will push the system from this neighborhood, and these excursions will be nearly iid (as in renewal theory). The problem of the exit time is then one of estimating the time till the first "successful" burst of noise, i.e., the one that leads to escape. The most likely way that will happen can be found by examining the quasipotential for the point on the boundary that is easiest to reach, i.e., the point (or points) that satisfies $Q(y) = \inf \{Q(z) : z \in \partial G\}$.

Since the probability of this easiest type of escape is $p \approx e^{-n \inf \{Q(z) : z \in \partial G\}}$, the distribution of the exit time should be Poisson with mean $1/p$, i.e., $E_{x_0} \tau^n \approx e^{n \inf \{Q(z) : z \in \partial G\}}$. If there is only one location y on ∂G where the

minimum is achieved, then the likelihood of escape from any other point should be smaller by an exponential factor, and thus the distribution of the exit location should concentrate near y as $n \to \infty$.

2 Comparison with Central Limit Theorem and moderate deviations

It is instructive to compare the information provided by a LDP with what one can get from the Central Limit Theorem (CLT) or one of its process-level generalizations These two limit theorems answer different questions, and provide complementary information. To simplify the discussion, consider for now the sum of n iid random variables. Here the CLT tells us something about probabilities for sets located distance $n^{1/2}$ from the mean, whereas the LDP considers sets that are much further into the tail (sets distance n from the mean). Suppose, for example, that $EX_1 = 0$, and that the moment generating function is finite for all α. (Of course this implies a finite second moment, and hence the CLT is also valid.) To simplify the notation assume the variables are normalized so that $EX_1^2 = 1$.

Consider a set A such that $P\{\theta \in \partial A\} = 0$, where θ is $N(0,1)$. Then by the CLT

$$P\left\{\frac{S_n}{n^{1/2}} \in A\right\} \to P\{\theta \in A\},$$

and so

$$P\left\{S_n \in n^{1/2}A\right\} \approx P\{\theta \in A\},$$

where $\approx$ means that the ratio tends to 1 as $n \to \infty$. Note that the CLT can tell us nothing about an estimate of the form $P\{S_n \in nA\}$, because this would require a limit for

$$P\left\{\frac{S_n}{n^{1/2}} \in n^{1/2}A\right\},$$

which is not valid since the set $n^{1/2}A$ depends on n.

In contrast, the LDP tells us about $P\{S_n \in nA\}$. Suppose that L is the rate function, and in analogy with $P\{\theta \in \partial A\} = 0$ assume that A is an L-continuity set with

$$C \doteq \inf_{\beta \in A^\circ} L(\beta) = \inf_{\beta \in \bar{A}} L(\beta).$$

Then

$$P\{S_n \in nA\} = P\left\{\frac{S_n}{n} \in A\right\} \approx e^{-nC},$$

in the sense that the ratio of the logarithms converge to 1. The LDP does not directly give us any CLT information, since the set A here is also fixed, and a CLT type statement would require

$$P\left\{\frac{S_n}{n} \in \frac{1}{n^{1/2}}A\right\}.$$

It does turn out that there is some CLT information hidden in the rate function, in that the local expansion up to order 2 of the rate function around its minimum point is determined by just means and variances, and is the same as what one would expect from a formal CLT approximation. This is related to the fact that *moderate deviations* provide a bridge between large deviations and the CLT, i.e., it gives limits for scalings between $n^{1/2}A$ and nA. Such a result is particularly useful for the design of accelerated Monte Carlo if the event of interest is far enough in the tail that a moderately large number of samples is required for good accuracy, but the cost per sample is very high (e.g., if the process model is a stochastic partial differential equation).

We will give the statement and proof of moderate deviations for the same diffusion model as in Lecture 4. This is a little misleading, in that the proof of the moderate deviations principle (MDP) for other process models is not such a straightforward adaptation of the proof of the corresponding LDP. In particular, for other processes (e.g., the process model of Lecture 6), because of the scaling it is a little harder to prove the tightness needed for the MDP.

Thus we focus on the process model

$$dX^\varepsilon = b(X^\varepsilon)dt + \sqrt{\varepsilon}\sigma(X^\varepsilon)dW, \quad X^\varepsilon(0) = x_0. \tag{8}$$

The same conditions as in Lecture 4 are assumed, and in addition that b is at least once continuously differentiable. Let

$$Y^\varepsilon(t) = \sqrt{\frac{a(\varepsilon)}{\varepsilon}}\left[X^\varepsilon(t) - X^0(t)\right],$$

where X^0 is the solution to (8) when $\varepsilon = 0$. We are interested in the LDP for Y^ε, when $a(\varepsilon) \to 0$ and $\varepsilon/a(\varepsilon) \to 0$.

Applying the representation gives

$$-a(\varepsilon)\log Ee^{-\frac{1}{a(\varepsilon)}F(Y^\varepsilon)} = \inf_{v\in\mathcal{U}} E\left[F(\bar{Y}^\varepsilon) + \frac{1}{2}\int_0^1 \|v(s)\|^2\,ds\right],$$

where

$$d\bar{X}^\varepsilon = b(\bar{X}^\varepsilon)dt + \sqrt{\varepsilon}\sigma(\bar{X}^\varepsilon)dW + \sqrt{\varepsilon/a(\varepsilon)}\sigma(\bar{X}^\varepsilon)vdt, \quad \bar{X}^\varepsilon(0) = x_0,$$

and

$$\bar{Y}^\varepsilon(t) = \sqrt{\frac{a(\varepsilon)}{\varepsilon}}\left[\bar{X}^\varepsilon(t) - X^0(t)\right].$$

2.1 Tightness and identification of limits

The issues of tightness and relating the limits are essentially the same as for the LDP. We can write

$$\bar{Y}^\varepsilon(t) = \bar{Y}_1^\varepsilon(t) + \bar{Y}_2^\varepsilon(t) + \bar{Y}_3^\varepsilon(t),$$

where

$$\begin{aligned}
\bar{Y}_1^\varepsilon(t) &= \sqrt{\frac{a(\varepsilon)}{\varepsilon}} \int_0^t \left[b(\bar{X}^\varepsilon(s)) - b(X^0(s)) \right] ds \\
\bar{Y}_2^\varepsilon(t) &= \sqrt{a(\varepsilon)} \int_0^t \sigma(\bar{X}^\varepsilon(s)) dW(s) \\
\bar{Y}_3^\varepsilon(t) &= \int_0^t \sigma(\bar{X}^\varepsilon(s)) v(s) ds.
\end{aligned}$$

As in the proof of the LDP we can assume that any sequence of interest $\{v^\varepsilon\}$ takes values in the compact set S_M for suitably large $M < \infty$, and therefore $\left\{(\bar{X}^\varepsilon, \bar{Y}_1^\varepsilon, \bar{Y}_2^\varepsilon, \bar{Y}_3^\varepsilon, v^\varepsilon), \varepsilon \in (0,1)\right\}$ is tight. Let ε index a convergent subsequence. It follows directly that $\bar{X}^\varepsilon \to X^0$ and $\bar{Y}_2^\varepsilon \to 0$. Assume that $\bar{Y}^\varepsilon \to \bar{Y}$ and $v^\varepsilon \to v$. Then

$$\bar{Y}_3^\varepsilon \to \int_0^\cdot \sigma(X^0(s)) v(s) ds.$$

Using that

$$\bar{X}^\varepsilon(t) = X^0(t) + \sqrt{\frac{\varepsilon}{a(\varepsilon)}} \bar{Y}^\varepsilon(t) \text{ and } \bar{Y}_1^\varepsilon \to \int_0^\cdot Db(X^0(s)) \bar{Y}(s) ds,$$

We conclude that

$$\bar{Y} = \int_0^\cdot \left[Db(X^0(s)) \bar{Y}(s) + \sigma(X^0(s)) v(s) \right] ds.$$

□

2.2 Statement of the MDP

The proof is now the same as that of the LDP. Thus Y^ε satisfies the LDP on $\mathcal{C}([0,1] : \mathbb{R}^d)$ with the rate

$$J(\phi) = \inf \left\{ \frac{1}{2} \int_0^1 \|v(s)\|^2 ds : \phi(t) = \int_0^t \left[Db(X^0(s)) \phi(s) + \sigma(X^0(s)) v(s) \right] ds \right\}$$

and scaling function $a(\varepsilon)$, which is the MDP for X^ε. Note that the only information that appears in the rate function for the MDP is a local expansion of the rate function I for the LDP around the zero cost path X^0. Note also that for suitable initial or terminal conditions (e.g., quadratic in $\phi(1)$) the infimization problem is a special case of the famous *linear-quadratic-regulator* from the theory of deterministic optimal control, and hence has an explicit solution.

Lecture 6: Processes That Are Not Functionals of an IID Noise Process

1 A Representation for Markovian Noise

For both Sanov's Theorem and the LDP for small noise diffusions, the "high-level" relative entropy representation could be expressed in a much simpler form. This happened because the quantities of interest could be represented as measurable functionals of a "white noise" process. However, such a representation is not always useful, even though it might be possible. For example, the underlying noise model might be Markovian, and though an "white noise" representation can be shown to exist, it might be hard to work with when proving, e.g., a law of large numbers limits.

In this lecture we will consider a "small noise" Markov process model. The particular model occurs frequently in stochastic systems theory, e.g., stochastic approximation and related recursive algorithms, and can be viewed as a substantial generalization of the random walk considered in Cramér's Theorem. It also arises as a discrete time approximation to various continuous time models, such as the SDE model of Lecture 4, and indeed provides an alternative approach to proving large deviation estimates for such models (though we much prefer the direct approach of Lecture 4).

2 Process Model

We begin with a description of the process model. Suppose that $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Thus $\theta(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$ for every $x \in \mathbb{R}^d$, and for every $A \in \mathcal{B}(\mathbb{R}^d)$ the mapping $x \to \theta(A|x)$ is Borel measurable. Then one can construct a probability space that supports iid random vector fields $\{v_i : \mathbb{R}^d \to \mathbb{R}^d, i \in \mathbb{N}_0\}$, with the property that for any $x \in \mathbb{R}^d$ $v_i(x)$ has distribution $\theta(\cdot|x)$. We then define for each $n \in \mathbb{N}$ a Markov process $\{X_i^n, i = 1, \ldots, n\}$ by setting

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0. \tag{9}$$

This discrete time process is interpolated into continuous time exactly as was done in Lecture 1, i.e.,

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right]. \tag{10}$$

Example 1 Suppose that for each $x \in \mathbb{R}^d$ $v_i(x)$ has a normal distribution with mean $b(x)$ and covariance $\sigma(x)\sigma^T(x)$. Then $X^n(t)$ can be viewed as the Euler approximation to the SDE of Lecture 4.

Example 2 For an example in the form of a stochastic approximation algorithm, one can take $v_i(x) = -\nabla V(x) + w_i$, where the w_i are iid with $Ew_i = 0$ and V is a smooth function. In this case $1/n$ is the "gain" of the algorithm.

To prove an LDP for $\{X^n, n \in \mathbb{N}\}$ additional assumptions must be made. For $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$ define

$$H(x, \alpha) \doteq \log E e^{\langle \alpha, v_i(x) \rangle}.$$

Condition 1 *We assume the following:*

1. *for each* $\alpha \in \mathbb{R}^d$ $\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty$,
2. *the mapping* $x \to \theta(\cdot|x)$ *from* $\mathbb{R}^d$ *to* $\mathcal{P}(\mathbb{R}^d)$ *is continuous in the topology of weak convergence,*
3. *for each* $x \in \mathbb{R}^d$ *the convex hull of the support of* $\theta(\cdot|x)$ *is* $\mathbb{R}^d$.

The first condition is needed for the rate function to have compact level sets. The second and third conditions can be weakened. With weakenings of the third condition the qualitative form of the large deviation rate function is the same, but additional "Lipschitz type" conditions on $x \to \theta(\cdot|x)$ are required. Such weakenings are quite important in applications (e.g., the third condition rules out bounded noise), but the proof involves a mollification and is considerably more complicated than the one we will give here. Those who are interested can find suitable conditions that weaken the third requirement in [14].

Weakening the second condition generally leads to a qualitatively different form of the rate function, and the process models that violate this condition are said to have "discontinuous statistics." Example 4 of Lecture 1 was of this form.

3 Representation

The first issue to resolve is the formulation of a representation that reflects the natural structure of the process model. Note that one could represent $\{X^n, n \in \mathbb{N}\}$ in terms of iid random variables, e.g., in the form $X^n_{i+1} = X^n_i + \frac{1}{n} G(X^n_i, U_i)$, where G is measurable and the $\{U_i, i \in \mathbb{N}_0\}$ are iid uniform. However, this form, which would allow a representation in terms of an iid reference measure, would not be useful. The map G would not in general be continuous, and this formulation would be poorly suited to proving even a law of large numbers limit.

An alternative and more useful representation follows from the form (9) and the chain rule for relative entropy. As before we will consider the

representation one would use to prove an LDP on $C([0,1]:\mathbb{R}^d)$, but $[0,1]$ could be replaced by any interval $[0,T]$, $T<\infty$. The line of argument will be to adapt the proof used for Sanov and Cramér to this functional setting. To obtain "process-level" information requires a more complicated empirical measure than was needed previously. Define L^n by

$$L^n(A\times B)\doteq\int_B L^n(A|t)dt,\quad L^n(A|t)=\delta_{v_i(X_i^n)}(A)\text{ if }t\in[i/n-1/n,i/n)$$

for Borel sets $A\subset\mathbb{R}^d$ and $B\subset[0,1]$. This measure (and its controlled analogue to be introduced) record the joint empirical distribution of velocity and time.

Theorem 1 *Let $G:\mathcal{P}(\mathbb{R}^d\times[0,1])\to\mathbb{R}$ be bounded and continuous and let L^n be defined as in the last display. Then*

$$-\frac{1}{n}\log Ee^{-nG(L^n)}=\inf_{\{\bar\mu_i^n\}}E\left[G(\bar L^n)+\frac{1}{n}\sum_{i=1}^n R\left(\bar\mu_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar X_i^n)\right)\right].\quad(11)$$

The infimum is over all controls $\{\bar\mu_i^n\}$, and the controls and controlled process $\bar X^n$ and measure $\bar L^n$ are recursively constructed as follows. Let $\bar X_0^n=x_0$ and define $\bar{\mathcal{F}}_i^n=\sigma(\bar X_j^n,j=0,\dots,i)$. Then $\bar\mu_i^n$ is measurable with respect to $\bar{\mathcal{F}}_i^n$ and gives the conditional distribution of $\bar v_i^n$. We set

$$\bar X_{i+1}^n=\bar X_i^n+\frac{1}{n}\bar v_i^n,$$

and then repeat. When $\left\{\bar X_i^n,i=1,\dots,n\right\}$ has been constructed, $\bar X^n(t)$ is defined as in (10) as the piecewise linear interpolation, and

$$\bar L^n(A\times B)\doteq\int_B\bar L^n(A|t)dt,\quad\bar L^n(A|t)=\delta_{\bar v_i^n}(A)\ \textit{if}\ t\in[i/n-1/n,i/n).$$

Note that the definition of $\bar L^n$ allows us to write

$$\bar X^n(t)=\int_{\mathbb{R}^d\times[0,t]}y\bar L^n(dy\times dt)+x_0.$$

Thus as in passing from Sanov to Cràmer, convergence of $\bar L^n$ plus some uniform integrability will imply convergence of $\bar X^n$.

Proof. The representation follows directly from the high-level variational representation for exponential integrals (Lemma 4 in Lecture 2) and the chain rule, and is essentially the same as what was done in deriving the representation used to prove Sanov's Theorem. The only difference is that the base measure there was product measure, reflecting the iid noise structure. Here the base measure is the Markov measure

$$\theta(dv_0|X_0^n)\cdot\theta(dv_1|X_1^n)\cdot\dots\cdot\theta(dv_{n-1}|X_{n-1}^n),$$

where

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i.$$

This change in the base measure is reflected by a change in the relative entropy cost, i.e., the $\bar{X}_i^n$ dependence in $\theta(\cdot|\bar{X}_i^n)$, which was just $\theta(\cdot)$ in the iid case. $\square$

Note that, as in the representation developed previously, we suppress the explicit dependence of $\bar{\mu}_i^n$ on $\bar{X}_j^n, j = 0, \ldots, i$, and account for this dependence by considering any $\bar{\mathcal{F}}_i^n$-measurable controls.

Before going further, we pause to comment on the expected form of the rate function. There is a time scale separation, which is due to the $1/n$ scaling and the weak continuity of $x \to \theta(\cdot|x)$. Over an interval $[s, s+\delta]$, with $\delta > 0$ small, the noise terms in the definition of $X^n(s+\delta) - X^n(s)$ are approximately iid $\theta(\cdot|X^n(s))$, and therefore by Cramér's Theorem

$$\frac{X^n(s+\delta) - X^n(s)}{\delta} \approx \frac{1}{n\delta} \sum_{i=\lfloor ns \rfloor}^{\lfloor ns+n\delta \rfloor} v_i(X^n(s))$$

will satisfy the LDP with rate $\delta L(X^n(s), \beta)$, where

$$L(x, \beta) = \inf \left\{ R\left(\mu(\cdot) \,\|\theta(\cdot|x)\right) : \int_{\mathbb{R}^d} y\mu(dy) = \beta \right\}. \tag{12}$$

Using the Markov property to combine estimates over small intervals and the heuristic LD approximation mentioned in Lecture 1, for a smooth trajectory ϕ that starts at x_0 and small $\sigma > 0$, an even more heuristic calculation gives

$$\begin{aligned}
&P\left\{X^n \in B_\sigma(\phi)\right\} \\
&\quad \approx \quad P\left\{X^n(j\delta) \in B_\sigma(\phi(j\delta)) \text{ all } 0 \le j \le \lfloor 1/\delta \rfloor\right\} \\
&\quad \approx \quad P\left\{ \frac{X^n(j\delta+\delta) - X^n(j\delta)}{\delta} \in B_\sigma\left(\frac{\phi(j\delta+\delta) - \phi(j\delta)}{\delta} \right),\ 0 \le j \le \lfloor 1/\delta \rfloor \right\} \\
&\quad \approx \quad \prod_{j=0}^{\lfloor 1/\delta \rfloor} e^{-n\delta L(\phi(j\delta), [\phi(j\delta+\delta)-\phi(j\delta)]/\delta)} \\
&\quad \approx \quad e^{-n\int_0^1 L(\phi(s), \dot{\phi}(s))ds},
\end{aligned}$$

and so one may expect the rate function $I(\phi) = \int_0^1 L(\phi(s), \dot{\phi}(s))ds$ for such ϕ.

We now turn to the rigorous analysis. As in the previous cases, we will establish tightness (and as with Cramér's Theorem a uniform integrability), and then prove a result which links the limits of weakly converging controls and controlled processes. We then prove the Laplace principle by showing a lower bound (which implies the large deviation upper bound) and an upper bound (which gives the large deviation lower bound). Besides the empirical

measures $\left\{\bar{L}^n, n \in \mathbb{N}\right\}$ we will make use of an analogous measure that records the time dependence of the $\bar{\mu}_i^n$, and therefore define

$$\bar{\mu}^n(A \times B) \doteq \int_B \bar{\mu}^n(A|t)dt, \quad \bar{\mu}^n(A|t) = \bar{\mu}_i^n(A) \text{ if } t \in [i/n - 1/n, i/n). \quad (13)$$

4 Statement and Proof of the LDP

Theorem 2 *Assume Condition 1, and define X^n by (10). Let*

$$I(\phi) = \begin{cases} \int_0^1 L(\phi(s), \dot{\phi}(s))ds & \text{if } \phi \in \mathcal{A}_{x_0}([0,1] : \mathbb{R}^d), \\ \infty & \text{otherwise,} \end{cases}$$

where $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is defined by (12). Then $\{X^n, n \in \mathbb{N}\}$ satisfies the LDP in $C([0,1] : \mathbb{R}^d)$ with rate I.

Note that the alternative expression

$$L(x, \beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(x, \alpha)]$$

mentioned when discussing Cramér's Theorem applies here as well. A proof of this fact will be given in Lemma 5. For the rest of this lecture we assume Condition 1.

4.1 Tightness and uniform integrability

Lemma 3 *Consider any sequence of controls $\{\bar{\mu}_i^n\}$ for which the relative entropy costs*

$$E\left[\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot) \,\middle\|\, \theta(\cdot|\bar{X}_i^n)\right)\right]$$

appearing in the representation (11) are uniformly bounded by $K < \infty$. Then the empirical measures $\left\{\bar{L}^n, n \in \mathbb{N}\right\}$ are tight and in fact uniformly integrable in the sense

$$\lim_{M \to \infty} \limsup_{n \to \infty} E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt)\right] = 0.$$

Also, the processes $\left\{\bar{X}^n, n \in \mathbb{N}\right\}$ are tight, as are the random measures $\{\bar{\mu}^n, n \in \mathbb{N}\}$.

Proof. Except for more complicated notation, the proof is almost the same as for Cramér's Theorem. We recall from Lecture 3 that if $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies $\mu \ll \theta$, then for any $\sigma \geq 1$

$$\int_{\mathbb{R}^d} \|y\| \, 1_{\{\|y\| \geq M\}} \mu(dy) \leq \int_{\mathbb{R}^d} e^{\sigma \|y\|} 1_{\{\|y\| \geq M\}} \theta(dy) + \frac{1}{\sigma} R\left(\mu \,\|\, \theta\right).$$

This is applied for each $i = 1, \ldots, n$ with $(\mu, \theta) = \left(\bar{\mu}_i^n(\cdot), \theta(\cdot|\bar{X}_i^n)\right)$:

$$\begin{aligned} & E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt)\right] \\ \leq & \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} e^{\sigma\|y\|} 1_{\{\|y\| \geq M\}} \theta(dy|x) + \frac{1}{\sigma} K. \end{aligned}$$

We claim that with σ fixed the first term vanishes as $M \to \infty$. Using the bound

$$e^{\sigma\|y\|} \leq \sum_{j=1}^{d} \left[e^{d\sigma y_j} + e^{-d\sigma y_j}\right],$$

it follows that for each j and k and choice of $\pm$,

$$\begin{aligned} \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} e^{\pm d\sigma y_j} 1_{\{\pm y_k \geq M\}} \theta(dy|x) & \leq e^{-\sigma M} \int_{\mathbb{R}^d} e^{\pm d\sigma y_j} e^{\pm \sigma y_k} \theta(dy|x) \\ & \leq e^{-\sigma M} e^{\sup_{x \in \mathbb{R}^d} H(x, \pm d\sigma e_j \pm \sigma e_k)} \\ & \to 0 \end{aligned}$$

and the claim follows from this. Sending first $M \to \infty$ and then $\sigma \to \infty$ gives tightness and uniform integrability of $\left\{\bar{L}^n, n \in \mathbb{N}\right\}$.

To establish tightness of $\left\{\bar{X}^n, n \in \mathbb{N}\right\}$ we use the fact that

$$\bar{X}^n(t) = \int_{\mathbb{R}^d \times [0,t]} y \bar{L}^n(dy \times dt) + x_0. \tag{14}$$

Tightness will follow if given $\varepsilon > 0$ there is $\delta > 0$ such that

$$\limsup_{n \to \infty} P\left\{w^n(\delta) \geq \varepsilon\right\} = 0, \tag{15}$$

where $w^n(\delta) \doteq \sup_{0 \leq s < t \leq 1 : t-s \leq \delta} \|X^n(t) - X^n(s)\|$. Let $\eta > 0$ be given, and choose $M < \infty$ such that

$$\limsup_{n \to \infty} E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt)\right] \leq \frac{\varepsilon\eta}{2}.$$

Let $\delta \doteq (\varepsilon/2M) \wedge 1$. Then since $M\delta \leq \varepsilon/2$,

$$\sup_{0 \leq s < t \leq 1 : t-s \leq \delta} \int_{\mathbb{R}^d \times [s,t]} \|y\| \, 1_{\{\|y\| \leq M\}} \bar{L}^n(dy \times dt) \leq M\delta \leq \frac{\varepsilon}{2}.$$

Hence

$$\begin{aligned} P\left\{w^n(\delta) \geq \varepsilon\right\} & \leq P\left\{\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt) \geq \frac{\varepsilon}{2}\right\} \\ & \leq \frac{2}{\varepsilon} E\left[\int_{\mathbb{R}^d \times [0,1]} \|y\| \, 1_{\{\|y\| \geq M\}} \bar{L}^n(dy \times dt)\right] \\ & \leq \eta. \end{aligned}$$

Since $\eta > 0$ is arbitrary, the limit (15) follows.

Finally, $\{\bar{\mu}^n, n \in \mathbb{N}\}$ is tight because each $\bar{\mu}^n$ is just a conditional mean of the corresponding $\bar{L}^n$, since $\bar{v}_i^n$ has conditional distribution $\bar{\mu}_i^n$ (see Lemma 9 of Lecture 2). □

4.2 Weak convergence

When taking limits we will need to keep track of $\theta(\cdot|\bar{X}_i^n)$. With this in mind we introduce

$$\lambda^n(A \times B) \doteq \int_B \lambda^n(A|t)dt, \quad \lambda^n(A|t) = \theta(A|\bar{X}_i^n) \text{ if } t \in [i/n - 1/n, i/n).$$

Lemma 4 *Consider any sequence of controls $\{\bar{\mu}_i^n\}$ for which the relative entropy costs*

$$E\left[\frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}_i^n(\cdot) \,\|\theta(\cdot|\bar{X}_i^n)\right)\right]$$

appearing in the representation (11) are uniformly bounded. Let $\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n), n \in \mathbb{N}\}$ denote a weakly converging subsequence, which for notational convenience we label by n, with limit $(\bar{X}, \bar{L}, \bar{\mu})$. Then w.p.1 $\bar{L} = \bar{\mu}$, and $\bar{\mu}(dy \times dt)$ can be decomposed as $\bar{\mu}(dy|t)dt$, where $\bar{\mu}(dy|t)$ is a stochastic kernel on $\mathbb{R}^d$ given $[0,1]$, and

$$\bar{X}(t) = \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy \times dt) + x_0 = \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy|t)dt + x_0. \tag{16}$$

In addition, λ^n converges weakly to a limit λ of the form

$$\lambda(A \times B) = \int_B \theta(A|\bar{X}(t))dt.$$

Proof. Precisely the same martingale argument as in the proof of Sanov's Theorem can be used to show that $\bar{L} = \bar{\mu}$ w.p.1. The uniform integrability allows us to pass to the limit in (14) and obtain

$$\bar{X}(t) = \int_{\mathbb{R}^d \times [0,t]} y\bar{L}(dy \times dt) + x_0.$$

Now use that $\bar{L} = \bar{\mu}$ w.p.1. to get the first part of (16). Since each $\bar{\mu}^n(dy \times dt)$ has Lebesgue measure as its second marginal the same is true for $\bar{\mu}$, and so both the decomposition and the second part of (16) follow. Finally, the weak convergence of λ^n and the form of the limit follow from the weak convergence of $\bar{X}^n$ to $\bar{X}$ and the assumption that $x \to \theta(A|x)$ is continuous. □

4.3 Lower bound

To prove the lower bound on the representation we again follow the line of argument used for Cramér's Theorem. The first step is to specialize the representation for application to $\{X^n, n \in \mathbb{N}\}$. Let $F : C([0,1] : \mathbb{R}^d) \to \mathbb{R}$ be bounded and continuous. Since X^n is a measurable function of L^n, the representation (11) implies

$$-\frac{1}{n}\log Ee^{-nF(X^n)} = \inf_{\{\bar{\mu}_i^n\}} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\big\|\theta(\cdot|\bar{X}_i^n)\right)\right].$$

The definitions of $\bar{\mu}^n$ and λ^n as measures that are piecewise constant in t gives the first inequality and the chain rule gives the second in the following:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\big\|\theta(\cdot|\bar{X}_i^n)\right) &= \int_{[0,1]} R\left(\bar{\mu}^n(\cdot|t)\,\|\lambda^n(\cdot|t)\right)dt \qquad (17)\\ &= R\left(\bar{\mu}^n(dy\times dt)\,\|\lambda^n(dy\times dt)\right).\end{aligned}$$

Now choose any $\varepsilon > 0$, and let $\{\bar{\mu}_i^n, i \in \{1,\dots,n\}\}$ satisfy

$$-\frac{1}{n}\log Ee^{-nF(X^n)} + \varepsilon \geq E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\big\|\theta(\cdot|\bar{X}_i^n)\right)\right].$$

Consider any subsequence of $\left\{(\bar{X}^n, \bar{\mu}^n, \lambda^n), n \in \mathbb{N}\right\}$ that converges to a weak limit $(\bar{X}, \bar{\mu}, \lambda)$, and denote the convergent subsequence by n. If the lower bound is demonstrated for this subsequence, the standard argument by contradiction establishes the lower bound for the original sequence. The details of the following calculation are given after the display:

$$\begin{aligned}&\liminf_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(X^n)} + \varepsilon\\ &\quad\geq \liminf_{n\to\infty} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\big\|\theta(\cdot|\bar{X}_i^n)\right)\right]\\ &\quad= \liminf_{n\to\infty} E\left[F(\bar{X}^n) + R\left(\bar{\mu}^n(dy\times dt)\,\|\lambda^n(dy\times dt)\right)\right]\\ &\quad\geq E\left[F\left(\bar{X}\right) + R\left(\bar{\mu}(dy\times dt)\,\|\lambda(dy\times dt)\right)\right]\\ &\quad= E\left[F\left(\bar{X}\right) + \int_{[0,1]} R\left(\bar{\mu}(dy|t)\,\big\|\theta(dy|\bar{X}(t))\right)dt\right]\\ &\quad\geq E\left[F\left(\bar{X}\right) + \int_{[0,1]} L(\bar{X}(t),\dot{\bar{X}}(t))dt\right]\\ &\quad\geq \inf_{\phi\in C([0,1]:\mathbb{R}^d)}\left\{F(\phi) + I(\phi)\right\}.\end{aligned}$$

The first equality uses the rewriting of the relative entropy in (17); the next inequality is due to the weak convergence, the lower semicontinuity

of $R(\cdot\,\|\cdot)$ and continuity of F, and Fatou's lemma; the next equality uses the decompositions $\bar{\mu}(dy \times dt) = \bar{\mu}(dy|t)dt$ and $\lambda(dy \times dt) = \theta(dy|\bar{X}(t))dt$; the third inequality uses Lemma 4 and the definition of L in (12); the last equality uses the definition of I and the fact that $\bar{X}(0) = x_0$ w.p.1.

4.4 I is a rate function

In the weak convergence approach, a deterministic version of the argument used to prove the lower bound usually shows that the proposed rate function is indeed a rate function, i.e., that it has compact level sets. Suppose that $\{\phi_j, j \in \mathbb{N}\}$ is given such that $I(\phi_j) \leq K < \infty$. Then we need to show $\{\phi_j, j \in \mathbb{N}\}$ is precompact, and that if $\phi_j \to \phi$ then

$$\liminf_{j\to\infty} I(\phi_j) \geq I(\phi).$$

It follows from the definition of L in (12) that for each j there is a probability measure $\mu^j(dy \times dt)$ such that

$$I(\phi_j) + \frac{1}{j} \geq \int_{[0,1]} R\left(\mu^j(dy|t)\,\big\|\theta(dy|\phi_j(t))\right) dt,$$

where $\mu^j(dy \times dt) = \mu^j(dy|t)dt$ and $\int_{\mathbb{R}^d} y\mu^j(dy|t) = \dot{\phi}_j(t)$. Using $I(\phi_j) \leq K < \infty$, exactly the same argument as in Lemma 3 shows that $\{\mu^j, j \in \mathbb{N}\}$ is tight and uniformly integrable, and a deterministic version of Lemma 4 shows that if μ^j denotes a convergent subsequence with limit μ, then $\phi_j \to \phi$ where $\int_{\mathbb{R}^d} y\mu(dy|t) = \dot{\phi}(t)$. Using the lower semicontinuity of relative entropy and the same arguments just used for the lower bound, we have

$$\begin{aligned}
\liminf_{j\to\infty}\left(I(\phi_j) + \frac{1}{j}\right) &\geq \liminf_{j\to\infty} R\left(\mu^j(dy \times dt)\,\big\|\theta(dy|\phi_j(t))dt\right) \\
&\geq \int_{[0,1]} R\left(\mu(dy|t)\,\|\theta(dy|\phi(t))\right) dt \\
&\geq \int_{[0,1]} L(\phi(t), \dot{\phi}(t))dt \\
&= I(\phi).
\end{aligned}$$

4.5 Upper bound

To prove the upper bound and complete the proof of the theorem, we must take a trajectory ϕ that nearly minimizes in $\inf_{\phi\in C([0,1]:\mathbb{R}^d)}\{F(\phi) + I(\phi)\}$ and show how to construct a control that can be applied to the representation which will have asymptotically the same cost. This requires some regularity properties of $L(x, \beta)$, which we now state. The proof of this result is given at the end of the notes.

Lemma 5 *Define $L(x,\beta)$ by (12). Then L is continuous on $\mathbb{R}^d \times \mathbb{R}^d$, and convex in β for each fixed x.*

We can now approximate a nearly minimizing trajectory and construct a control for the representation. Fix $\varepsilon > 0$. Then there is $\psi \in C([0,1] : \mathbb{R}^d)$ such that $[F(\psi) + I(\psi)] \leq \inf_{\phi \in C([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)] + \varepsilon$. While $\{\psi(t) : 0 \leq t \leq 1\}$ is bounded by continuity, we also claim we can assume, without loss of generality, that $\{\dot{\psi}(t) : 0 \leq t \leq 1\}$ is bounded. This claim will be shown after we complete the proof.

Let $M < \infty$ and $K < \infty$ be such that

$$\sup_{t\in[0,1]} \|\psi(t)\| \vee \sup_{t\in[0,1]} ||\dot{\psi}(t)|| \leq M, \qquad \sup_{(x,\beta):\|x\|\leq M+1, \|\beta\|\leq M+1} L(x,\beta) \leq K.$$

For $\delta > 0$ let ψ^δ be the piecewise linear interpolation of ψ, with interpolation points $t = k\delta$. Since $\sup_{t\in[0,1]} ||\psi^\delta(t)|| \leq M$ and $\sup_{t\in[0,1]} ||\dot{\psi}^\delta(t)|| \leq M$, the dominated convergence theorem implies there is $\delta > 0$ such that $\left[F(\psi^\delta) + I(\psi^\delta)\right] \leq [F(\psi) + I(\psi)] + \varepsilon$. We set $\phi^* = \psi^\delta$ for such a δ.

The construction of a control to apply in the representation is now straightforward. Given (x,β), one can choose $\mu^*_{x,\beta}$ in a measurable way (as a function of x,β) such that

$$\int_{\mathbb{R}^d} y\mu^*_{x,\beta}(dy) = \beta \text{ and } R\left(\mu^*_{x,\beta}(dy)\,\|\theta(dy|x)\right) \leq L(x,\beta) + \varepsilon.$$

Recall that $\bar{\mu}^n_i$ depends on time and is also allowed to be any measurable function of $\bar{X}^n_j, j = 0,\ldots,i$. Define $N^n = \inf\{j : ||\bar{X}^n_j - \phi^*(j/n)|| > 1\} \wedge n$. Then we set

$$\bar{\mu}^n_i(\cdot) = \begin{cases} \mu^*_{\bar{X}^n_i, \dot{\phi}^*(i/n)}(\cdot) & \text{if } i \leq N^n \\ \theta(\cdot|\bar{X}^n_i) & \text{if } i > N^n \end{cases}.$$

The cost under this control satisfies

$$E\left[\frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}^n_i(\cdot)\,\big\|\theta(\cdot|\bar{X}^n_i)\right)\right] \leq E\left[\frac{1}{n}\sum_{i=1}^{N^n} L(\bar{X}^n_i, \dot{\phi}^*(i/n)) + \varepsilon\right] \leq K + \varepsilon.$$

It follows that Lemma 3 applies. Since $\tau^n \doteq N^n/n$ takes values in a compact set, given any subsequence of $\mathbb{N}$ we can find a further subsequence (again denoted by n) such that $(\bar{X}^n, \bar{\mu}^n, \tau^n)$ converges in distribution to a limit $(\bar{X}, \bar{\mu}, \tau)$. It follows from the fact that the mean of $\bar{\mu}^n_i$ is $\dot{\phi}^*(i/n)$ for $i \leq N^n$ that

$$\int_{\mathbb{R}^d\times[0,t]} y\bar{\mu}(dy|t)dt + x_0 = \phi^*(t)$$

for all $t \in [0,\tau]$, and therefore $\bar{X}(t) = \phi^*(t)$ for all $t \in [0,\tau]$, w.p.1. The definition of N^n implies that whenever $\tau < 1$, $\limsup_{s\downarrow\tau} \left\|\bar{X}(s) - \phi^*(\tau)\right\| \geq 1$. However this contradicts the fact that $\bar{X}$ has continuous sample paths,

and thus $\tau = 1$ w.p.1. Therefore along the full sequence $\mathbb{N}$, $\bar{X}^n$ converges in distribution to ϕ^*.

We can now put the pieces together to prove the upper bound. For the particular control $\{\bar{\mu}_i^n\}$ just constructed, we have

$$\begin{aligned}
&\limsup_{n\in\mathbb{N}} -\frac{1}{n}\log Ee^{-nF(X^n)} \\
&\leq \limsup_{n\in\mathbb{N}} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right)\right] \\
&\leq \limsup_{n\in\mathbb{N}} E\left[F(\bar{X}^n) + \frac{1}{n}\sum_{i=1}^n L(\bar{X}_i^n, \dot{\phi}^*(i/n)) + \varepsilon\right] \\
&= \left[F(\phi^*) + \int_{[0,1]} L(\phi^*(t), \dot{\phi}^*(t))dt + \varepsilon\right] \\
&\leq \inf_{\phi\in C([0,1]:\mathbb{R}^d)} [F(\phi) + I(\phi)] + 3\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the upper bound (and hence the large deviation lower bound) follows.

Lemma 6 *Consider $\psi \in C([0,1]:\mathbb{R}^d)$ such that $[F(\psi) + I(\psi)] < \infty$. Then given $\varepsilon > 0$, there is ψ^* such that $\{\dot{\psi}^*(t) : 0 \leq t \leq 1\}$ is bounded and $[F(\psi^*) + I(\psi^*)] \leq [F(\psi) + I(\psi)] + \varepsilon$.*

Proof. For $\lambda \in (0,1)$ let $D_\lambda \doteq \{t : ||\dot{\psi}(t)|| \geq 1/\lambda\}$, and define a time rescaling $S_\lambda : [0,1] \to [0,\infty)$ by $S_\lambda(0) = 0$ and

$$\dot{S}_\lambda(t) = \begin{cases} ||\dot{\psi}(t)||/(1-\lambda) & t \in D_\lambda \\ 1 & \text{otherwise.} \end{cases}$$

Then $S_\lambda(t)$ is continuous and strictly increasing. Let T_λ be defined by $T_\lambda(S_\lambda(t)) = t$. Then T_λ is defined on $[0, S_\lambda(1)] \supset [0,1]$, and when considered on $[0,1]$

$$\psi_\lambda(t) \doteq \psi(T_\lambda(t))$$

is a "slowed" version of ψ. By the chain rule $\dot{\psi}_\lambda(S_\lambda(t)) = \dot{\psi}(t)/\dot{S}_\lambda(t)$, and therefore $\dot{\psi}_\lambda(t)$ has uniformly bounded derivative. ψ^* in the lemma will be ψ_λ for large but finite λ.

First note that since $I(\psi) < \infty$ we have

$$\lim_{\lambda\to 0}\int_0^1 1_{D_\lambda}(t)L(\psi(t),\dot{\psi}(t))dt = 0.$$

Owing to part 1 of Condition 1, $L(x,\beta)$ is uniformly superlinear in β, i.e., $L(x,\beta)/||\beta|| \to \infty$ uniformly in x as $||\beta|| \to \infty$. Therefore the last display implies

$$\lim_{\lambda\to 0}\int_0^1 1_{D_\lambda}(s)||\dot{\psi}(t)||dt = 0,$$

which in turn implies $\lim_{\lambda\to 0} S_\lambda(s) = s$ uniformly for $s \in [0,1]$. Since

$$\sup_{t\in[0,1]} ||\psi_\lambda(t)-\psi(t)|| = \sup_{t\in[0,1]} ||\psi(T_\lambda(t))-\psi(t)|| = \sup_{t\in[0,T_\lambda(1)]} ||\psi(s)-\psi(S_\lambda(s))||,$$

$\sup_{t\in[0,1]} ||\psi_\lambda(t) - \psi(t)|| \to 0$ as $\lambda \to 0$.

Thus we need only show that $I(\psi_\lambda)$ is close to $I(\psi)$. Let

$$\Gamma \doteq \sup_{t\in[0,1]} \sup_{\beta:||\beta||\le 1} L(\psi(t),\beta) < \infty.$$

For $t \in D_\lambda$ the non-negativity of L implies

$$\begin{aligned} L\left(\psi(t), \frac{\dot\psi(t)}{\dot S_\lambda(t)}\right)\dot S_\lambda(t) - L(\psi(t),\dot\psi(t)) &\le L\left(\psi(t), \frac{(1-\lambda)\dot\psi(t)}{||\dot\psi(t)||}\right)\frac{||\dot\psi(t)||}{1-\lambda} \\ &\le \frac{\Gamma}{1-\lambda}||\dot\psi(t)|| \end{aligned}$$

and therefore

$$\begin{aligned} I(\psi_\lambda) - I(\psi) &\le \int_0^{S_\lambda(1)} L(\psi_\lambda(t),\dot\psi_\lambda(t))dt - \int_0^1 L(\psi(t),\dot\psi(t))dt \\ &= \int_0^1 L(\psi_\lambda(S_\lambda(t)),\dot\psi_\lambda(S_\lambda(t)))\dot S_\lambda(t)dt - \int_0^1 L(\psi(t),\dot\psi(t))dt \\ &= \int_0^1 L\left(\psi(t), \frac{\dot\psi(t)}{\dot S_\lambda(t)}\right)\dot S_\lambda(t)dt - \int_0^1 L(\psi(t),\dot\psi(t))dt \\ &\le \frac{\Gamma}{1-\lambda}\int_0^1 1_{D_\lambda}(s)||\dot\psi(t)||dt. \end{aligned}$$

We can now let $\lambda \to 0$. □

4.6 Proof of Lemma 5

Proof. We first show that L is finite on $\mathbb{R}^d \times \mathbb{R}^d$. Since for each x the support of $\theta(\cdot|x)$ is all of $\mathbb{R}^d$ (part 3 of Condition 1), the map $\alpha \to H(x,\alpha)$ is superlinear, which implies that for each $x \in \mathbb{R}^d$ the gradient $D_\alpha H(x,\alpha)$ is onto $\mathbb{R}^d$. Since the map is also convex, given β there is a unique vector $\alpha(\beta)$ such that $D_\alpha H(x,\alpha(\beta)) = \beta$. Setting $\mu(dy) = e^{\langle\alpha(\beta),y\rangle}\theta(dy|x)/e^{H(x,\alpha(\beta))}$, we have

$$\int_{\mathbb{R}^d} y\mu(dy) = \frac{1}{e^{H(x,\alpha(\beta))}}\int_{\mathbb{R}^d} ye^{\langle\alpha(\beta),y\rangle}\theta(dy|x) = D_\alpha H(x,\alpha(\beta)) = \beta.$$

Direct calculation using the form of $\mu(dy)$ and the definition of relative entropy then gives

$$L(x,\beta) \le R\left(\mu(dy)\,\|\theta(dy|x)\right) = \langle\alpha(\beta),\beta\rangle - H(x,\alpha(\beta)) < \infty. \tag{18}$$

The convexity of L in β follows directly from the definition of L. Suppose μ_i, $i = 1, 2$, come within $\varepsilon > 0$ of achieving the infimum for β_i, $i = 1, 2$. Then for any $\rho \in [0, 1]$

$$\int_{\mathbb{R}^d} y \left[\rho\mu_1(dy) + (1-\rho)\mu_2(dy)\right] = \rho\beta_1 + (1-\rho)\beta_2.$$

Since relative entropy is convex,

$$\begin{aligned} L(x,\beta) &\leq R\left(\left[\rho\mu_1(dy) + (1-\rho)\mu_2(dy)\right] \| \theta(dy|x)\right) \\ &\leq \rho L(x,\beta_1) + (1-\rho)L(x,\beta_2) + \varepsilon, \end{aligned}$$

and the convexity holds since $\varepsilon > 0$ is arbitrary. Using the lower semicontinuity of relative entropy, one can likewise show that $L(x, \beta)$ is jointly lower semicontinuous, and hence $\beta \to L(x, \beta)$ is a proper convex function for each fixed x.

We next claim that

$$L(x,\beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta\rangle - H(x,\alpha)\right]. \tag{19}$$

Using the definition of L, we have

$$H(x,\alpha) = \sup_{\mu \in \mathcal{P}(\mathbb{R}^d)} \left[\int_{\mathbb{R}^d} \langle \alpha, y\rangle \, \mu(dy) - R\left(\mu(dy) \| \theta(dy|x)\right)\right].$$

Consider the sequence $g_{N,\alpha}(y) = [\langle \alpha, y\rangle \vee -N] \wedge N$ of bounded continuous functions. By duality for relative entropy (Lemma 4 of Lecture 2)

$$\log \int_{\mathbb{R}^d} e^{g_{N,\alpha}(y)} \theta(dy|x) = \sup_{\mu \in \mathcal{P}(\mathbb{R}^d)} \left[\int_{\mathbb{R}^d} g_{N,\alpha}(y)\mu(dy) - R\left(\mu(dy) \| \theta(dy|x)\right)\right].$$

As discussed previously $H(x, \alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$ implies $\int_{\mathbb{R}^d} e^{\sigma||y||}\theta(dy|x) < \infty$ for all $\sigma \in (0, \infty)$, which provides a dominating function to use when taking limits on the left hand side. Now fix any measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ for which $R\left(\mu(dy) \| \theta(dy|x)\right) < \infty$. As was argued in the proof of Cramér's Theorem, we have uniform integrability of $g_{N,\alpha}$. Passing to the limit in

$$\log \int_{\mathbb{R}^d} e^{g_{N,\alpha}(y)} \theta(dy|x) \geq \int_{\mathbb{R}^d} g_{N,\alpha}(y)\mu(dy) - R\left(\mu(dy) \| \theta(dy|x)\right)$$

gives

$$R\left(\mu(dy) \| \theta(dy|x)\right) \geq \left\langle \alpha, \int_{\mathbb{R}^d} y\mu(dy) \right\rangle - H(x,\alpha),$$

and therefore for all α

$$L(x,\beta) \geq \langle \alpha, \beta\rangle - H(x,\alpha).$$

By (18) there is an α for which the reverse inequality holds, and therefore (19) follows.

Joint continuity of $H(x, \alpha)$ follows from the weak continuity of $x \to \theta(dy|x)$ and the dominated convergence theorem. We now show that joint continuity of $L(x, \beta)$ follows from this. If a sequence of differentiable convex functions g_i with Legendre transforms g_i^* converge pointwise to another differentiable convex function g with transform g^*, and if β is any point such that $g^*(\beta) < \infty$, then whenever $\beta_i \to \beta$ we have $g_i^*(\beta_i) \to g^*(\beta)$ [14, Lemma C.8.1]. We apply this result with $g_i^*(\alpha) = H(x_i, \alpha)$ and $g(\alpha) = H(x, \alpha)$, to conclude that if $x_i \to x$ and $\beta_i \to \beta$, then $L(x_i, \beta_i) \to L(x, \beta)$. □

Lecture 7: Extracting Information From the Variational Problem

In this lecture we remark on methods for solving the variational problem that arises in the large deviation study of time dependent, nearly deterministic systems. As one might suspect there is no single technique–one needs to use whatever structure the problem offers. Methods for approximating variational problems associated with empirical measure problems are less well studied, and a few remarks on this topic are given in Lecture 11.

1 Example problems

To illustrate the different methods but at the same time keep the discussion focused, we will consider a few canonical problems. There are many similar and closely related problems which allow an analogous treatment. Throughout, we assume there is a process $\{X^\varepsilon, \varepsilon \in (0,1)\}$ for which a large deviation principle holds for any initial condition $x \in \mathbb{R}^d$ and any interval $[0,T]$. We assume that the rate function takes the form

$$I_T(\phi) = \int_0^T L(\phi(t), \dot{\phi}(t))dt$$

whenever it is finite. When the process is suitably "stationary," this is very mild assumption. Additional assumptions will be placed on L in the sequel, but at a minimum it is assumed non-negative and lower semicontinuous, and convex in the second variable. Note that $L(x,\beta) = \infty$ is allowed. This happens when the noise in the system does not push the state in all directions in $\mathbb{R}^d$.

Example 1 (Hitting a rare set before entering a neighborhood of a stable point) We consider two Borel sets $A, B \subset \mathbb{R}^d$, where A contains a global attractor for the zero cost trajectories of the rate function, and an initial condition $x \in (A \cup B)^c$. The problem of interest is the estimation of

$$P_x \left\{ X^\varepsilon \text{ enters } B \text{ before entering } A \right\}.$$

One can prove under appropriate conditions that

$$-\varepsilon \log P_x \left\{ X^\varepsilon \text{ enters } B \text{ before entering } A \right\} \to V(x),$$

where

$$V(x) \doteq \inf \left\{ \int_0^T L(\phi(t), \dot{\phi}(t))dt : \phi \in C_{x,T},\, T < \infty \right\},$$

and

$$C_{x,T} = \{\phi(0) = x, \phi(t) \in B \text{ for some } t \in [0,T] \text{ and } \phi(s) \notin A \text{ for } s \in [0,t]\}.$$

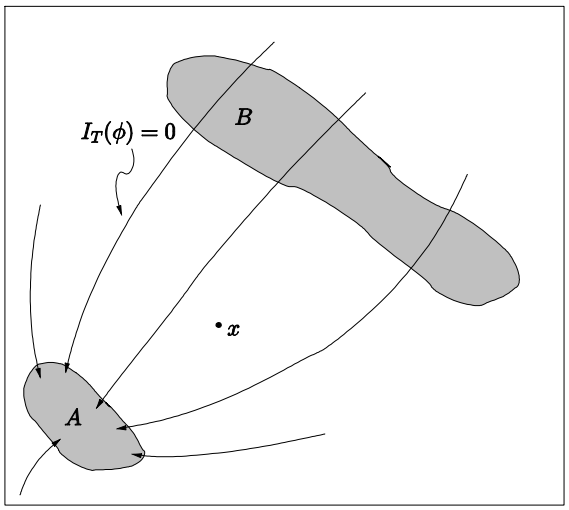

Figure 8: Stability of the zero cost trajectories

There are two types of conditions (besides the large deviation principle) needed for such a limit to hold. The first is a regularity condition on A, B and L which essentially guarantees that the infimum of the rate over the interior and closure of $C_{x,T}$ agree. The second is an assumption on $\Lambda(x) \doteq \{\beta : L(x,\beta) = 0\}$ which asserts that all solutions to $\dot{\phi}(t) \in \Lambda(\phi(t))$ converge in some sense to A as $t \to \infty$. The multi-dimensional random walk (Example 1 of Lecture 1) is an example of this type, although in this case A is not needed (or could be interpreted as points "at $-\infty$"). Such a criterion would also be suitable for the diffusion process that models a tracking loop of Lecture 1 (with A a neighborhood of 0 and B the complement of the domain of attraction), as well as the queueing model of Lecture 1 (with $A = \{0\}$ and B corresponding to the second queue exceeding the buffer size).

Example 2 (Problems on a finite time interval) We consider a Borel set $B \subset \mathbb{R}^d$ and $T \in (0,\infty)$, and use large deviations to estimate either

$$P_x\{X^\varepsilon \text{ enters } B \text{ before } T\} \text{ or } P_x\{X^\varepsilon(T) \in B\}.$$

In the first case we assume any solution to $\dot{\phi}(t) \in \Lambda(\phi(t)), \phi(0) = x$ will not enter B by T, and in the second case $\phi(T) \notin B$. As before, regularity conditions on ∂B and L are required for the infimum over the closure and interior to agree. Under these conditions we have

$$-\varepsilon \log P_x\{X^\varepsilon \text{ enters } B \text{ before } T\} \to V(x),$$

where

$$V(x) \doteq \inf\left\{\int_0^T L(\phi(t),\dot{\phi}(t))dt : \phi(0) = x, \phi(t) \notin B \text{ for all } t \in [0,T]\right\},$$

and similarly for the second case.

Example 3 (RISK-SENSITIVE COST) Let $F : \mathbb{R}^d \to \mathbb{R}$ be bounded and continuous and let $T \in (0, \infty)$, and suppose we seek to approximate $E_x e^{-\frac{1}{\varepsilon}F(X^\varepsilon(T))}$. Then

$$-\varepsilon \log E_x e^{-\frac{1}{\varepsilon}F(X^\varepsilon(T))} \to V(x),$$

where

$$V(x) \doteq \inf \left\{ \int_0^T L(\phi(t), \dot{\phi}(t))dt + F(\phi(T)) : \phi(0) = x \right\}.$$

Depending on the use, one may want to solve for just a particular initial condition, find the solution for all initial conditions (i.e., the solution to the related PDE), or something in between (e.g., subsolutions of the PDE–see Lectures 9 and 10).

2 Related PDE

The PDE related to these examples can be derived by a formal application of dynamic programming. Here we will just state the PDE and relevant terminal/boundary conditions. The use of these PDE will also be formal, and in particular we will not worry about questions of uniqueness.

PDE for Example 1 The PDE for the first example is as follows. Let

$$H(x, \alpha) = \sup_{\beta \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - L(x, \beta) \right],$$

and recall that in the large deviation theory developed in previous lectures H had an interpretation as a log-moment generating function or some similar quantity, and was often available in explicit form. Let $DV(x)$ denote the gradient of $V : \mathbb{R}^d \to \mathbb{R}$. The PDE based on dynamic programming that should characterize $V(x)$ is

$$\inf_{\beta \in \mathbb{R}^d} \left[\langle DV(x), \beta \rangle + L(x, \beta) \right] = 0 \text{ for } x \in (A \cup B)^c$$

together with the boundary conditions

$$V(x) = 0 \text{ for } x \in \partial B \text{ and } V(x) = \infty \text{ for } x \in \partial A.$$

In the appropriate framework of viscosity solutions the condition on ∂A is relaxed to $V(x) \leq \infty$, and therefore may be ignored. It is convenient to introduce the function $\mathbb{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$\mathbb{H}(x, p) = \inf_{\beta \in \mathbb{R}^d} \left[\langle p, \beta \rangle + L(x, \beta) \right] = - \sup_{\beta \in \mathbb{R}^d} \left[\langle -p, \beta \rangle - L(x, \beta) \right] = -H(x, -p)$$

so that the PDE is written as $\mathbb{H}(x, DV(x)) = 0$ for $x \in (A \cup B)^c$. Note that since $H(x, \alpha)$ is convex in α, $\mathbb{H}(x, p)$ is concave in p.

PDE for Example 2 For the second example, we should first extend the definition of V to allow any initial condition $(t, x), t \in [0, T], x \notin B$, and minimize over all trajectories that satisfy the initial condition $\phi(t) = x$. The PDE for the problem of entering a rare set B during an interval $[0, T]$ takes the form

$$V_t(t, x) + \mathbb{H}(x, DV(t, x)) = 0 \text{ for } (t, x) \in (0, T) \times B^c,$$

together with the boundary and terminal conditions

$$V(t, x) = 0 \text{ for } (t, x) \in [0, T) \times \partial B \text{ and } V(T, x) = \infty \text{ for } x \in B^c.$$

The solution to the variational problem corresponds to $V(0, x)$.

PDE for Example 3 The PDE for the problem involving the risk-sensitive cost $E_x e^{-\frac{1}{\varepsilon}F(X^\varepsilon(T))}$ takes the form

$$V_t(t, x) + \mathbb{H}(x, DV(t, x)) = 0 \text{ for } (t, x) \in (0, T) \times \mathbb{R}^d, \tag{20}$$

together with the terminal condition $V(T, x) = F(x)$.

3 Using convexity

The simplest problems for which a more-or-less explicit solution is possible are those which exploit convexity.

Example 4 We consider again the multi-dimensional random walk model (Example 1 of Lecture 1), and the associated variational problem. To simplify the notation let $d = 2$. Assume b is the unique point for which $L(b) = 0$, and that $b_i < 0$ for $i = 1, 2$. The variational problem can then be reduced to

$$\begin{aligned} V(x) &\doteq \inf\left\{\int_0^T L(\dot\phi(t))dt : \phi(0) = x, \phi_1(T) \geq M_1 \text{ or } \phi_2(T) \geq M_2, T < \infty\right\} \\ &= V_1(x) \wedge V_2(x), \end{aligned}$$

where

$$V_i(x) = \inf\left\{\int_0^T L(\dot\phi(t))dt : \phi(0) = x, \phi_i(T) \geq M_i, T < \infty\right\}, i = 1, 2.$$

Since L is convex by Jensen's inequality $\int_0^T L(\dot\phi(t))dt \geq TL([\phi(T) - x]/T)$, and hence one can restrict to paths with constant velocity in $V_i(x)$, i.e.,

$\phi(t) = x + t\beta$. To satisfy the constraint requires $\beta_i \geq [M_i - x_i]/T$, and it is easy to argue that the minimizer will satisfy this with equality. Thus

$$V_i(x) = \inf \{TL(\beta) : T\beta_i = (M_i - x_i), T < \infty\}.$$

It is easy to check that $V_i(x)$ is affine in x, with a gradient in the direction of e_i (the standard basis vectors). Thus $V_i(x) = \langle x, p^i \rangle + c_i$, with $p^i_j = 0$ if $i \neq j$.

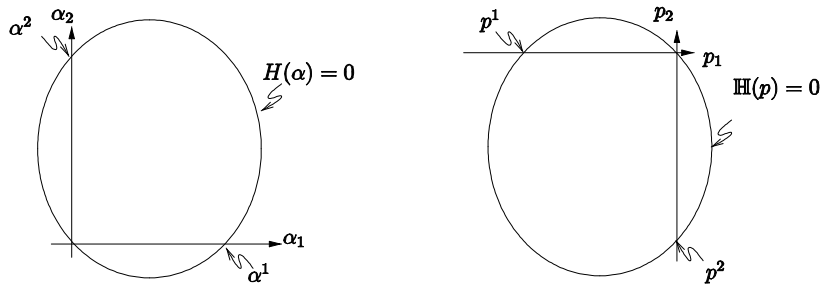

Figure 9: Roots for $\mathbb{H}(p) = 0$

Since V_i should satisfy $\mathbb{H}(DV_i(x)) = 0$ we find $\mathbb{H}(p^i) = 0$, and using the boundary condition to solve for c_i gives $V_i(x) = \langle (x - M_i e_i), p^i \rangle$, where the roots p^i are illustrated in Figure 9. That the relevant roots of $H(\alpha) = 0$ have positive components is due to the facts that $H(0) = 0, DH(0) = b < 0$ and convexity, while the negativity of the roots to $\mathbb{H}(p) = 0$ follows from $\mathbb{H}(p) = -H(-p)$. Having constructed a candidate solution, one can use a verification argument (with a suitable generalization of the chain rule) to prove that in fact $V(x) = V_1(x) \wedge V_2(x)$.

Remark 1 Many problems from stochastic networks have "sectionally homogeneous" dynamics, such as the queueing example (Example 4 in Lecture 1), and as a consequence $L(x, \beta)$ has a very structured dependence on x. In Example 4, $L(x, \beta)$ is independent of x in any of the sets S, D, ∂_1 and ∂_2 indicated in Figure 10. Using convexity of $L(x, \cdot)$ within these regions, one can argue that minimizing trajectories have constant velocities within these sets, and thereby reduce the minimization problem over trajectories to a finite dimensional minimization. Examples of this sort can be found in [22, 21, 25, 29, 30].

4 Other explicit solutions

There are other classes of system dynamics for which one can obtain an explicit solution. The most well known of these corresponds to large deviations

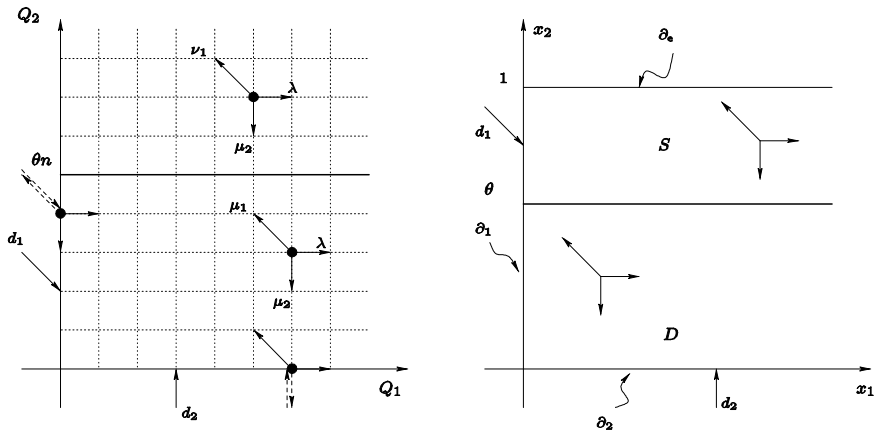

Figure 10: Jump rates and partition of the state space for the scaled system.

for the linear SDE

$$dX^\varepsilon(t) = \bar{B}(t)X^\varepsilon(t)dt + \sqrt{\varepsilon}\bar{A}(t)dW(t),$$

where $\bar{B}(t)$ is deterministic and takes values in the space of $d \times d$ matrices, $\bar{A}(t)$ takes values in the space of $d \times k$ matrices, and $W(t)$ is a k-dimensional standard Brownian motion. Assume that $\bar{B}(t)$ and $\bar{A}(t)$ are continuous. It follows from Lecture 4 that the local rate function $L(x, \beta)$ for this process takes the form

$$L(x,\beta) = \frac{1}{2}\left\langle \left(\beta - \bar{B}(t)x\right), \left[\bar{A}(t)\bar{A}(t)^T\right]^{-1}\left(\beta - \bar{B}(t)x\right)\right\rangle \qquad (21)$$

if $\bar{A}(t)\bar{A}(t)^T$ is positive definite. When $\bar{A}(t)\bar{A}(t)^T$ is only non-negative definite there is a similar form in those directions β in which the noise can push the system, and value ∞ for all other directions.

Owing to the quadratic form of $L(x, \beta)$, explicit solutions are possible in terms of the famous *linear-quadratic-regulator* (LQR) from the theory of deterministic optimal control [46]. Consider the risk-sensitive cost in Example 3 with F quadratic. If one assumes the form

$$V(t,x) = \frac{1}{2}\left\langle (x - b(t)), a(t)\left(x - b(t)\right)\right\rangle + c(t)$$

and inserts this into (20) then a coupled system of ODEs for a, b and c is obtained, together with terminal conditions from $V(T, x) = F(x)$. From these one can construct a solution to the PDE, and again a verification argument can be used to show that $V(t, x)$ indeed characterizes the solution to the variational problem for all x and t.

The LQR can also be used to solve other problems that are not of exactly the LQR form, though only for particular initial conditions. For example,

consider the problem of approximating the probability $P_x\{X^\varepsilon(T) \in B\}$. This leads to the same PDE as the LQR, but with the terminal condition

$$F(x) = \begin{cases} 0 & x \in B \\ \infty & x \in B^c \end{cases} .$$

One can consider bounding F from below by the minimum of a finite collection of quadratic functions $U_j(x), j = 1, \ldots, J$, and then solving the LQR for each terminal condition U_j to produce $V_j(t,x)$. This yields the lower bound

$$V(t,x) \geq \min_{j=1,\ldots,J} V_j(t,x)$$

which under suitable conditions can be made tight by optimizing over the finite set of parameters that describe $U_j(x), j = 1, \ldots, J$. For example, if $d = 1$ and $B = [l_1, l_2]$ is an interval, then it suffices to consider $U_i(x) = b_i(x - l_i), i = 1, 2$ with $b_1 > 0$ and $b_2 < 0$ (see Lecture 9).

Although the form (21) appears special, we would note that this is the generic form that appears with moderate deviations (see Lecture 5), regardless of whether the underlying process model is Gaussian or not. Finally, we remark that other broad classes of models exist which admit explicit solutions but which are less well known. An example of such are the variational problems one needs to solve in connection with the occupancy problems (Example 5 of Lecture 1) [31].

5 Gradient systems

Recall that the quasipotential with respect to the starting point 0 was defined in Lecture 5 by

$$Q(x) = \inf\left\{\int_0^T L(\phi, \dot{\phi})dt : \phi(0) = 0, \phi(T) = x, T < \infty\right\}.$$

Although none of Examples 1-3 can be directly solved based on just the quasipotential, their solutions can sometimes be approximated in terms of Q, and there are other variational problems such as those mentioned in Lecture 5 for which it gives the exact solution. It was remarked in Example 1 of Lecture 5 that $Q(x)$ could be found explicitly for certain types of gradient systems, and here we will fill in some of the details using a classical verification argument. Assume $L(x, \beta) = (\beta - b(x))^2/2$, where 0 is a global attractor for b and that $b(x) = -DU(x) + \ell(x)$, where U is smooth with its unique local minimum of 0 at 0 and $\langle DU(x), \ell(x)\rangle = 0$. The claim is that Q has the explicit solution $2U$.

We will use the control form

$$Q(x) = \inf\left\{\int_0^T \frac{1}{2}\|u\|^2 dt : \dot{\phi} = b(\phi) + u, \phi(0) = 0, \phi(T) = x, T < \infty\right\}.$$

First note that a formal dynamic programming argument suggests that Q should satisfy

$$0 = \inf_{u\in\mathbb{R}^d} \left[-\langle DQ(x), b(x)+u\rangle + \frac{1}{2}\|u\|^2 \right] = \left[-\langle DQ(x), b(x)\rangle - \frac{1}{2}\|DQ(x)\|^2 \right].$$

Trying the proposed solution indeed gives

$$-2\langle DU(x), -DU(x)+\ell(x)\rangle - 2\|DU(x)\|^2 = 0$$

since by assumption $\langle DU(x), \ell(x)\rangle = 0$. From

$$\inf_{u\in\mathbb{R}^d} \left[-2\langle DU(x), b(x)+u\rangle + \frac{1}{2}\|u\|^2 \right] = 0$$

it follows that

$$-2\langle DU(x), b(x)+u\rangle + \frac{1}{2}\|u\|^2 \geq 0$$

for all $u \in \mathbb{R}^d$, with equality if and only if $u = 2DU(x)$.

Now consider any control u in the definition of Q. By the chain rule and the last display

$$\frac{d}{dt}2U(\phi(t)) = 2\langle DU(\phi(t)), b(\phi(t))+u(t)\rangle \leq \frac{1}{2}\|u(t)\|^2,$$

and integrating gives

$$\int_0^T \frac{1}{2}\|u(t)\|^2\, dt \geq 2U(x) - 2U(0) = 2U(x).$$

Thus $Q(x) \geq 2U(x)$. To prove the reverse inequality we solve $\dot{\phi} = b(\phi) + 2DU(\phi) = \ell(\phi) + DU(\phi)$ backward in time with terminal condition $\phi(T) = x$. This corresponds to using $u(t) = 2DU(\phi(t))$, and since equality holds gives

$$\int_0^T \frac{1}{2}\|u(t)\|^2\, dt = 2U(x) - 2U(\phi(0)).$$

Stability of b and the condition $\langle DU(x), \ell(x)\rangle = 0$ imply $\phi(0) \to 0$ as $T \to \infty$, and therefore $Q(x) \leq 2U(x)$.

Explicit formulas for the quasipotential are not limited to such Gaussian rate functions, and even hold for some infinite dimensional problems [39].

6 Numerical methods

When an analytic solution or approximate solution is not possible one can attempt a numerical solution. However, these are also limited in their range of application and/or effectiveness. There are methods which attempt a global

solution to the variational problem, either directly (e.g., Markov chain approximations as in Kushner-Dupuis or infinite dimensional linear programming as in Gaitsgory) or through the PDE (e.g., finite difference approximations as in Bardi-Falcone or the max-plus method of McEneaney). Other methods seek a solution to the variational problem at a particular initial condition. Examples in this category are the string method as in E-Vanden-Eijnden and the elastic band method of Jonsson, which utilize a descent in the space of trajectories to find a local minimizer, as well as shooting methods based on the Euler-Lagrange equations. This latter class of methods is better suited to problem with large dimension (if one is interested in only a particular initial condition), but as noted it may produce only a local minimizer to the variational problem.

Lecture 8: An Overview of Importance Sampling for Rare Events

In previous lectures we have described how large deviation theory gives approximations for various distributions, and in particular approximations to expected values and probabilities that are largely determined by rare events. These approximations take the form of logarithmic asymptotics, i.e., exponential decay rates.[2] For some purposes, especially when one is seeking qualitative information on how a rare event occurs, these approximations may be sufficient. However the quantitative value of the approximation may not be sufficient, and improved estimates may be sought.

In this situation it is natural to turn to Monte Carlo approximation. However, as we will explain in some detail, the Monte Carlo approximation of small probabilities and related expected values also has difficulties owing to the role of rare events, and the design of reliable schemes requires great care. It turns out that many of the tools and constructions used for the large deviation analysis of a given problem can be used for the problem of designing Monte Carlo schemes that are efficient and reliable. These topics will be discussed in this lecture and the two that follow.

1 Example of a quantity to be estimated

We return to a problem mentioned in Lecture 7, which is to evaluate

$$P_x \{X^n \text{ enters } B \text{ before entering } A\},$$

where the set A is an attractor of the noiseless system and B is rare. As a process model we consider the setup of Lecture 6, which is the general Markov model based on iid random vector fields $\{v_i(y), y \in \mathbb{R}^d\}$, and defined by

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x$$

and

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right].$$

We recall the notation

$$H(y, \alpha) = \log E \exp \langle \alpha, v_i(y) \rangle, \quad L(y, \beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(y, \alpha)].$$

We also recall that under some mild regularity conditions

$$-\frac{1}{n} \log P_x \{X^n \text{ enters } B \text{ before entering } A\} \to V(x),$$

[2] For certain special structures one can obtain more accurate approximations, e.g., approximations which identify both the exponential rate of decay as well as "pre-exponential" terms.

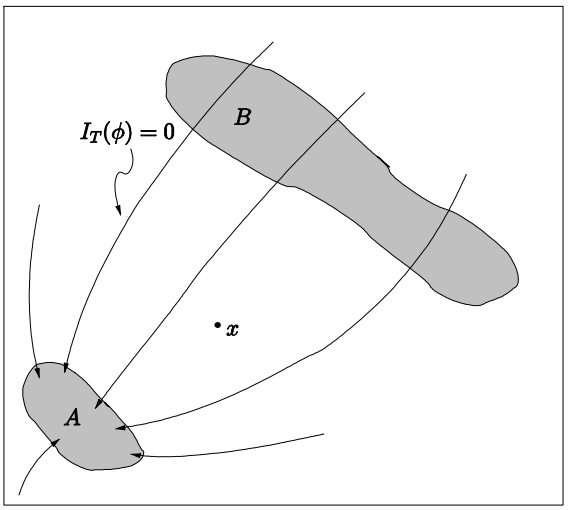

Figure 11: Stability of the zero cost trajectories

where

$$V(x) \doteq \inf\left\{\int_0^T L(\phi(t), \dot{\phi}(t))dt : \phi \in C_{x,T},\, T < \infty\right\},$$

and

$$C_{x,T} = \{\phi(0) = x, \phi(t) \in B \text{ for some } t \in [0,T] \text{ and } \phi(s) \notin A \text{ for } s \in [0,t]\}.$$

2 Basics of Monte Carlo

The problem of interest is to estimate

$$p^n(x) \doteq P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

Let $C_x \doteq \cup_{T\in(0,\infty)} C_{x,T}$, the trajectories that enter B before entering A. For standard Monte Carlo one simulates K independent copies $\{X_k^n, k = 1, \ldots, K\}$ of X^n, and then forms the estimate

$$\hat{p}_K^n(x) \doteq \frac{1}{K}\sum_{k=1}^{K} 1_{\{X_k^n \in C_x\}}.$$

Note that k here is the index of the sample and *not* the time step.

The variance of a single sample is

$$\begin{aligned}
\mathrm{Var}(1_{\{X_k^n \in C_x\}}) &= E\left[1_{\{X_k^n \in C_x\}} - E1_{\{X_k^n \in C_x\}}\right]^2 \\
&= E1_{\{X_k^n \in C_x\}} - \left(E1_{\{X_k^n \in C_x\}}\right)^2 \\
&= p^n(x) - (p^n(x))^2,
\end{aligned}$$

and since $p^n(x)$ is exponentially small $(p^n(x))^2$ can be neglected. The *relative error*, which is defined by the ratio of the standard deviation of $\hat{p}_K^n(x)$ and $p^n(x)$, is then

$$\frac{\sqrt{\mathrm{Var}(\hat{p}_K^n(x))}}{p^n(x)} \approx \sqrt{\frac{p^n(x)}{K}} \cdot \frac{1}{p^n(x)} = \sqrt{\frac{1}{Kp^n(x)}}.$$

Thus to obtain a relative error of size roughly 1 requires $K \approx e^{nV(x)}$ samples. This is computationally infeasible when $p^n(x)$ is very small (e.g., 10^{-5}), or even when $p^n(x)$ is not so small if the computational effort needed to generate samples of X^n is great. This is typical in many problems where rare event estimation is important. For example, consider the problem of estimating the probability of an unusually large concentration of pollutant in a model for ground water contamination. The generation of each sample would typically involve solving a time dependent partial differential equation, and hence each sample is computationally expensive.

An alternative is to construct iid random variables $\gamma_1^n, \ldots, \gamma_K^n$ with $E\gamma_1^n = p_n(x)$, and use the unbiased estimator

$$\hat{q}_K^n(x) \doteq \frac{\gamma_1^n + \cdots + \gamma_K^n}{K}.$$

The performance as with ordinary Monte Carlo is determined by variance of γ_1^n, and since $E\gamma_1^n = p_n(x)$, minimizing the variance is equivalent to minimizing $E\,(\gamma_1^n)^2$.

It is easy to obtain bounds on the best possible performance. For example, by Jensen's inequality

$$-\frac{1}{n}\log E\,(\gamma_1^n)^2 \leq -\frac{2}{n}\log E\gamma_1^n = -\frac{2}{n}\log p_n(x) \to 2V(x).$$

Hence the *decay rate* for the second moment cannot possible exceed $2V(x)$. An estimator is called *asymptotically efficient* if

$$\liminf_{n\to\infty} -\frac{1}{n}\log E\,(\gamma_1^n)^2 \geq \ 2V(x),$$

i.e., the decay rate is achieved.

One could consider more stringent measures of performance, such as *bounded relative error*: there is $K < \infty$ such that

$$\limsup_{n\to\infty} \frac{\sqrt{\mathrm{Var}(\gamma_1^n)}}{p^n(x)} \leq K,$$

and for systems with very simple structure this is sometimes possible. However, even mild complicating features (e.g., state dependent coefficients in a small noise diffusion) make this not feasible.

There are at least two well known methods to design random variables $\{\gamma_k^n\}$ that are unbiased, which can be simulated with reasonable effort, and for which one hopes to get good performance: *importance sampling* and *splitting schemes*. In these notes we will focus on importance sampling (IS), though many of the constructions needed for the successful design and analysis are essentially the same for both approaches [10, 11].

We stress that for any approach to this problem a rigorous and *independent* analysis of performance is very important, since typical methods one would use to assess accuracy of the estimates (e.g., the empirical variance) are prone to the same difficulties and errors which can affect the estimates themselves. This point will be illustrated in the numerical examples.

3 Importance sampling

3.1 Generalities

The idea of importance sampling is simple. Suppose the random variable X taking values in a Polish space S has distribution θ, and the goal is to approximate $M = Ef(X) = \int_S f(x)\theta(dx)$. An alternative probability measure μ on S is introduced with the property that θ is absolutely continuous with respect to μ, so the Radon-Nykodym derivative (or likelihood ratio) $[d\theta/d\mu](x)$ is well defined. One then simulates a random variable Y with distribution μ, and forms the estimate $f(Y) \cdot [d\theta/d\mu](Y)$, which is unbiased since

$$Ef(Y)\frac{d\theta}{d\mu}(Y) = \int_S f(x)\frac{d\theta}{d\mu}(x)\mu(dx) = \int_S f(x)\theta(dx).$$

The (single sample) variance of this estimator is then

$$E\left[f(Y)\frac{d\theta}{d\mu}(Y) - M\right]^2 = E\left[f(Y)\frac{d\theta}{d\mu}(Y)\right]^2 - M^2.$$

Depending on the choice of μ, the variance may be larger or smaller than ordinary Monte Carlo. The benefits (and dangers!) are especially acute in the rare event setting, since as we will see $[d\theta/d\mu]$ scales *exponentially* in n, the large deviation parameter.

Consider the case of estimating a probability, so that $1_{\{Y\in A\}}[d\theta/d\mu](Y)$ is used to estimate $p = P\{X \in A\}$. Ordinary Monte Carlo uses an average of 0's and 1's to approximate p. If p is very close to 0 (or 1) it will take many samples, since 0 and 1 are far away from each other (relative to p). A well designed IS scheme will cluster the samples near p through the design of $d\theta/d\mu$. However, in a complicated system the actual values taken by $d\theta/d\mu$ may be hard to predict, and they could be very far from p.

3.2 Importance sampling for rare events

Next we consider the particular problem of estimating $p^n(x)$. The first question is, "what are natural changes of measure?" A hint is provided by the analysis of Lecture 6. The control measures $\bar{\mu}_i^n$ of the weak convergence approach are equivalent to a new distribution for the noises under a change of measure. An a posteriori conclusion of the analysis is that *exponential* changes of measure are asymptotically optimal for the large deviations analysis. Exponential changes of measure have a finite dimensional parameterization, and thus are very convenient to work with. Recalling that $\{v_i(x), i \in \mathbb{N}\}$ are iid with distribution $\theta(dv|x)$ and associated log moment generating functions $H(x,\alpha)$, this suggests measures of the form

$$\nu_\alpha(dv|x) = e^{\langle \alpha, v\rangle - H(x,\alpha)}\theta(dv|x)$$

be used to generate the noise sequence under the new distribution. The parameter α can be thought of as a control to be selected to produce good performance of the resulting Monte Carlo scheme.

While more complicated dependencies could be considered, it will turn out that allowing α to depend on time and the current state of the simulated trajectory will be sufficient for asymptotic efficiency, and thus a control scheme (i.e., a change of measure) will be characterized as a collection of measurable mappings $\alpha_i^n(x)$, defined for $i \in \mathbb{N}_0$ and $x \in \mathbb{R}^d$, and taking values in $\mathbb{R}^d$. The generation of a single sample as well as the likelihood ratio needed to estimate $p^n(x)$ then proceeds as follows.

We initialize with $Y_0^n = x$. A sequence of noises w_i^n and states Y_{i+1}^n are then generated recursively by

$$P\left\{ w_i^n \in dv \middle| \mathcal{F}_i^n \right\} = \nu_{\alpha_i^n(Y_i^n)}(dv|Y_i^n), \text{ with } \mathcal{F}_i^n = \sigma\left(w_j^n, j = 0, \ldots, i\right)$$

and

$$Y_{i+1}^n = Y_i^n + \frac{1}{n} w_i^n.$$

The simulation proceeds up until

$$N^n = \inf\left\{ i : Y_i^n \in A \cup B \right\},$$

and we define $Y^n(t)$ to be the piecewise linear interpolation, so that $1_{\{Y^n \in C_x\}}$ means B was entered before A. The likelihood ration is then

$$\prod_{i=0}^{N^n-1} \frac{d\theta(\cdot|Y_i^n)}{d\nu_{\alpha_i^n(Y_i^n)}(\cdot|Y_i^n)}(w_i^n) = \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))},$$

and the estimate based on a single sample is thus

$$1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))}. \tag{22}$$

One then simulates K independent copies of (22) and takes the sample average.

We recall that performance is determined by the variance of a single sample, and minimizing this is the same as minimizing the second moment. The second moment of (22) is

$$E\left[1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-2\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + 2H(Y_i^n, \alpha_i^n(Y_i^n))}\right],$$

which when rewritten in terms of the distribution of the *original process* $\{X_i^n\}$ takes the form

$$E\left[1_{\{X^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(X_i^n), v_i(X_i^n) \right\rangle + H(X_i^n, \alpha_i^n(X_i^n))}\right].$$

4 A standard heuristic, and dangers in the rare event setting

Since one of the classical approaches to the large deviation lower bound involves a change of measure argument, it is natural to ask if there is a connection between the change of measure used there to prove bounds for a particular event or expected value, and the change of measure that might yield a good IS scheme for that same event. A few example problems suggested this to be true, and for some time it was generally thought that using the lower bound change of measure would work in general. This turned out to be false, and indeed the class of schemes that had been considered up to that time turned out to be, in general, inadequate. In this section we illustrate the issue through an example due to [36].

The example is as follows. Suppose that $v_i(X_i^n)$ are in fact independent of X_i^n, i.e., that they are just an iid sequence with distribution θ. We further assume $d = 1$ and that $x = 0$. Then X_i^n is a random walk, and $X_n^n = \sum_{i=0}^{n-1} v_i/n$ is just the sample mean, i.e., we are in the setting of Cramér's Theorem with rate function $L(\beta)$. Let $C \subset \mathbb{R}$, and suppose we want to estimate $P\{X_n \in C\}$ by importance sampling.

The heuristic just described to construct an alternative sampling distribution is straightforward. Let β^* solve $\inf\{L(\beta) : \beta \in C\}$ (and assume the infimum over the interior and closure of C are the same). If α^* is *dual* to β^*, i.e., if α^* is the point that maximizes in the relation $L(\beta^*) = \sup_{\alpha \in \mathbb{R}}[\alpha\beta^* - H(\alpha)]$, then as discussed in Lecture 6 the mean of $\nu_{\alpha^*}(dv) = e^{\langle \alpha^*, v \rangle - H(\alpha^*)}\theta(dv)$ is exactly β^*, and ν_{α^*} is the control we would use to prove the large deviation lower bound for a set that contains β^*. Since this problem is over a fixed time horizon the single sample estimate is just

$$1_{\{Y_n^n \in C\}} \prod_{i=0}^{n-1} e^{-\left\langle \alpha^*, w_i^n \right\rangle + H(\alpha^*)} = 1_{\{Y_n^n \in C\}} e^{-n[\langle \alpha^*, Y_n^n \rangle - H(\alpha^*)]}.$$

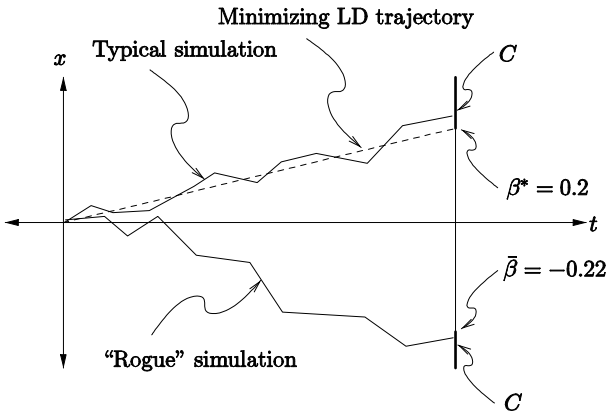

Figure 12: An expected trajectory and a rogue trajectory

We can now easily see the shortcomings of this heuristic. Assume that θ is Gaussian $N(0,1)$ and consider the *non-convex* set $C = (-\infty, -0.22] \cup [0.2, \infty)$. For this process $L(\beta) = \beta^2/2$, $H(\alpha) = \alpha^2/2$, and $\alpha^* = \beta^* = 0.2$, and the change of measure will shift the mean to this value. If all goes according to plan and the simulated trajectory ends up near β^*, then the likelihood ratio will be near $\exp -n\left[\langle \alpha^*, \beta^* \rangle - H(\alpha^*)\right] = \exp -nL(\beta^*)$, which is exactly what is needed. However, it is also possible that a rare event under the $\nu_{\alpha^*}(dv)$ distribution could occur, and one might end up with Y_n^n near $\bar{\beta}$. Such an occurrence is labeled the "rogue" trajectory in Figure 1. When this happens, the likelihood ration will be approximately

$$\exp -n\left[\langle \alpha^*, \bar{\beta} \rangle - H(\alpha^*)\right] = \exp n\left[0.2 \times 0.22 + \frac{1}{2}(0.2)^2\right].$$

This quantity grows exponentially in n and, while the event itself might be rare, it happens enough that the variance of the estimate is very large, and even larger than standard Monte Carlo!

In this example the true probability for $n = 100$ is $p_n = 3.67 \times 10^{-2}$. The following data reflects four trials of $K = 5000$ replications each.

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}_n$ ($\times 10^{-2}$)	2.23	2.24	17.32	16.37
Standard Error ($\times 10^{-2}$)	0.05	0.05	14.98	14.10
95% C.I. ($\times 10^{-2}$)	[2.13, 2.33]	[2.14, 2.34]	[-12.64, 47.28]	[-11.83, 44.57]

In the first two trials there were no "rogue" trajectories. The two estimates are smaller than the true value, and in fact neither of the confidence intervals contains the true value. Thus the same difficulties that affect

the estimation of p_n also make the confidence intervals essentially useless, though one does not a priori know this to be the case. Because of this, an independent theoretical (and not data driven) analysis of errors is essential for rare event Monte Carlo estimation. The second pair of trial include at least one rogue trajectory, which is needed to avoid the bias of the first two trials. The estimates are far from the true value, but in this case at least the confidence intervals are correctly indicating this fact.

One could argue that this example can be avoided by splitting the problem into estimating two half-infinite intervals. While such an approach would work here, it will fall apart as soon as one considers problems in higher dimensions or even slightly more complicated dynamics. What is needed is a *global* approach that properly controls the likelihood ratio for any possible simulated trajectory.

5 A dynamic game interpretation of importance sampling

Further insight into the difficulties of IS in the rare event setting can be obtained by modeling the performance with respect to the natural family of schemes in terms of *deterministic differential game*. In this section we formally develop this approximation for the simple random walk model just discussed (it actually holds quite broadly). Suppose that for the iid random walk model we consider, instead of the constant control α^* suggested by the standard heuristic, a collection of sampling controls of the form suggested previously, and in particular assume

$$\alpha_i^n(Y_i^n) = u(Y_i^n, i/n)$$

for some smooth function $u : \mathbb{R} \times [0,1] \to \mathbb{R}$. In this case, the second moment of a single sample, and hence the performance of the scheme, is given by the exponential integral

$$E\left[1_{\{Y_n^n \in C\}} \prod_{i=0}^{n-1} e^{-2u(Y_i^n, i/n)w_i^n + 2H(u(Y_i^n, i/n))}\right],$$

which we can rewrite in terms of the original process as

$$E\left[1_{\{X_n^n \in C\}} \prod_{i=0}^{n-1} e^{-u(X_i^n, i/n)v_i + H(u(X_i^n, i/n))}\right].$$

This is the same type of quantity we have considered in previous lectures.[3] It is expected to scale exponentially in n, and thus it is natural to

[3] Since the exponent is not bounded the results of [14] (Lemma 5 of Lecture 2) do not apply. However, under the assumed bound on the moment generating function the representation can be extended to cover the case where the exponent is affine in v_i.

consider the log transform and the corresponding relative entropy representation. Using the same notation for the controls (measures) and controlled processes as previously, we have

$$\begin{aligned}
&-\frac{1}{n}\log E\left[1_{\{X_n^n\in C\}}\prod_{i=0}^{n-1}e^{-u(X_i^n,i/n)v_i+H(u(X_i^n,i/n))}\right]\\
&= \inf_{\{\bar\mu_i^n\}} E\left[\frac{1}{n}\sum_{i=1}^{n}\left[u(\bar X_i^n,i/n)\bar v_i - H(u(\bar X_i^n,i/n))\right]\right.\\
&\qquad\left.+\frac{1}{n}\sum_{i=1}^{n}R\left(\bar\mu_i^n\,\|\theta\right)+\infty 1_{C^c}\left(\bar X_n^n\right).\right]
\end{aligned}$$

One can analyze the limit using weak convergence as was done previously. Under the LLN scaling we have (approximately and in a weak sense)

$$\bar X_i^n\approx\phi(i/n),\quad E\bar v_i\approx\dot\phi(i/n)\approx E\int_{\mathbb{R}}y\bar\mu_i^n(dy),\ \text{and}\ R\left(\bar\mu_i^n\,\|\theta\right)\approx L(\dot\phi(i/n)),$$

and so the limit produces the optimization problem

$$\begin{aligned}
&C[u]=\\
&\quad\inf_{\phi}\left[\int_{[0,1]}\left[u(\phi(t),t)\dot\phi(t)-H(u(\phi(t),t))+L(\dot\phi(t))\right]dt+\infty 1_{C^c}\left(\phi(1)\right)\right],
\end{aligned}$$

where the infimum is over absolutely continuous ϕ with $\phi(0)=0$.

The quantity $C[u]$ gives the rate of decay of the second moment of the IS scheme that uses the sampling control $\alpha_i^n(Y_i^n)=u(Y_i^n,i/n)$ to dynamically choose the change of measure. There are in fact two controls, which are $u(y,t)$, which is in feedback form but fixed for the analysis, and $\dot\phi$ [i.e., we are considering a calculus of variations form of the control problem]. This latter control arises in the large deviations analysis of the second moment. Since $C[u]$ is the rate of decay, the IS control u will want to make this as large as possible, and so one can consider the problem $U=\sup_{u(\cdot,\cdot)}C[u]$. This is a kind of deterministic differential (or dynamic) game, where $\dot\phi$ attempts to minimize (in open loop form) and u attempts to maximize (in feedback form, but being selected before ϕ is chosen).

We will not delve deeply into the nuances of differential games, since this game has a special structure which allows a reduction to a much simpler problem. We observe that the running cost takes the for $\alpha\beta-H(\alpha)+L(\beta)$ with $\alpha=u(\phi(t),t)$ and $\beta=\dot\phi(t)$. Note that this cost satisfies the min/max property:

$$\sup_{\alpha\in\mathbb{R}}\inf_{\beta\in\mathbb{R}}\left[\alpha\beta-H(\alpha)+L(\beta)\right]=\inf_{\beta\in\mathbb{R}}\sup_{\alpha\in\mathbb{R}}\left[\alpha\beta-H(\alpha)+L(\beta)\right]=2\inf_{\beta\in\mathbb{R}}L(\beta).$$

Suppose we extend the definition to allow for an arbitrary initial condition (x,t) (i.e., we consider the cost over $[t,1]$ and with $\phi(t)=x$), and denote the

corresponding optimal rate of decay by $U(x,t)$. Let U_t be the partial with respect to t and $DU(x,t)$ the gradient in x. Then $U(x,t)$ will be a viscosity solution to

$$U_t(x,t) + \sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} [DU(x,t)\beta + \alpha\beta - H(\alpha) + L(\beta)] = 0 \tag{23}$$

and the terminal condition

$$U(x,1) = \infty \text{ for } x \in C^c \text{ and } U(x,1) = 0 \text{ for } x \in C.$$

Note that the PDE can be rewritten as

$$0 = U_t(x,t) + \inf_{\beta\in\mathbb{R}} [DU(x,t)\beta + 2L(\beta)] = U_t(x,t) + 2\mathbb{H}(DU(x,t)/2).$$

Recalling from Lecture 7 that the large deviation rate $V(x,t)$ for such initial conditions will satisfy the same terminal condition and the PDE

$$V_t(x,t) + \mathbb{H}(DV(x,t)) = 0,$$

a comparison principle will imply $U(x,t) = 2V(x,t)$, which is the best possible rate of decay. The PDE for U (called an Isaacs equation in this context) also tells us (at least for smooth solutions) optimal controls for both players. Evaluating the infimum in β in (23) gives

$$\begin{aligned} & \sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} [DU(x,t)\beta + \alpha\beta - H(\alpha) + L(\beta)] \\ = & \sup_{\alpha\in\mathbb{R}} \left[- \sup_{\beta\in\mathbb{R}} [-(DU(x,t)+\alpha)\beta - L(\beta)] - H(\alpha) \right] \\ = & - \inf_{\alpha\in\mathbb{R}} [H(-DU(x,t)-\alpha) + H(\alpha)], \end{aligned}$$

and since H is convex the optimum is at $\alpha = -DU(x,t)/2$. In terms of the solution to the equation that governs the large deviation rate this is simply $\alpha = -DV(x,t)$.

Although this formal derivation was for the iid random walk model appearing in Cramér's Theorem, the same logic holds generally and for both the diffusion model of Lecture 4 and the state dependent random walk model of Lecture 6 one has the analogous PDEs and the analogous formula for the formally optimum state feedback sampling control, which is $\alpha_i^n(x) = -DV(x,t)$. For the problem of hitting a rare set B before A the PDE is stationary and with appropriate boundary conditions, but the change of measure suggested by the analysis takes the analogous form without time dependence: $\alpha^n(x) = -DV(x)$.

Lecture 9: The Subsolutions Approach to Importance Sampling

In the last lecture we discussed the shortcomings of the "open loop" sampling schemes that had been used previously for rare event sampling, and described how the introduction of feedback allowed (at least formally) for the existence of asymptotically optimal sampling controls. However, the constuction of the controls was in terms of the solution to a game (or as it turned out an equivalent control problem), and a rigorous analysis was not given.

It turns out that one does not need to solve the game or control problem, and in fact the construction of suitable *subsolutions* to the associated PDE will be sufficient. This is a significant simplification, because for many interesting classes of problems such subsolutions can be constructed explicitly. The reason subsolutions suffice is because one bound on performance (an upper bound on the rate of decay of the second moment) was automatic due to Jensen's inequality. To prove the reverse bound will require only certain inequalities, which coincide with the subsolution definition.

In the next section the definitions of classical and piecewise classical subsolution will be given. It will turn out to be much easier for many problems to find appropriate piecewise classical subsolutions, so this generalization is important. We also spell out how the various subsolutions generate sampling schemes.

1 Subsolutions

We will describe the subsolutions needed for both finite time problems and exit problems. We begin with the finite time problem, which generalizes the example used in Lecture 8. For convenience, we recall the notation and construction of the state dependent random walk model of Lecture 6. $\left\{v_i : \mathbb{R}^d \to \mathbb{R}^d, i \in \mathbb{N}_0\right\}$ are a collection of iid random vector fields with distribution $\theta(\cdot|x) \in \mathcal{P}(\mathbb{R}^d)$, i.e.,

$$P\left\{v_i(x) \in A\right\} = \theta(A|x).$$

We then define for each $n \in \mathbb{N}$ a Markov process $\{X_i^n, i = 1, \ldots, n\}$ by setting

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0,$$

and $X^n(t)$ is defined by piecewise linear interpolation:

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right].$$

Also, we assume $H(x,\alpha) = \log E \exp \langle \alpha, v_i(x) \rangle < \infty$ for all $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$.

The importance sampling problem of interest is to estimate

$$P_{x_0} \{ X^n(T) \in C \},$$

where $C \subset \mathbb{R}^d$. Recall that the PDE that characterizes the large deviation rate and half the optimal rate of decay for an asymptotically optimal importance sampling scheme is

$$V_t(x,t) + \mathbb{H}(x, DV(x,t)) = 0 \tag{24}$$

for $(x,t) \in \mathbb{R}^d \times [0,T)$, where $\mathbb{H}(x,p) = -H(x,-p)$. The terminal condition is

$$V(x,T) = \infty \text{ for } x \in C^c \text{ and } V(x,T) = 0 \text{ for } x \in C, \tag{25}$$

To simplify notation we assume Tn is an integer.

Definition 4 A function $\bar{V} : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ is a classical sense subsolution (or just classical subsolution) if it is continuously differentiable in both variables and if

$$\bar{V}_t(x,t) + \mathbb{H}(x, D\bar{V}(x,t)) \geq 0$$

for all $(x,t) \in \mathbb{R}^d \times [0,T)$ and if

$$\bar{V}(x,T) \leq \infty \text{ for } x \in C^c \text{ and } \bar{V}(x,T) \leq 0 \text{ for } x \in C.$$

Note that the condition $\bar{V}(x,T) \leq \infty$ for $x \in C^c$ is vacuous. Let $\wedge_{j=1}^J a_j$ denote the minimum of real numbers $a_j, j = 1, \ldots, J$.

Definition 5 A function $\bar{V} : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ is a piecewise classical sense subsolution (or just piecewise classical subsolution) if the following hold. There is $J \in \mathbb{N}$ and functions $\bar{V}^{(j)} : \mathbb{R}^d \times [0,T] \to \mathbb{R}, j = 1, \ldots, J$, that are continuously differentiable in both variables and satisfy

$$\bar{V}_t^{(j)}(x,t) + \mathbb{H}(x, D\bar{V}^{(j)}(x,t)) \geq 0$$

for all $(x,t) \in \mathbb{R}^d \times [0,T]$. Moreover $\bar{V}(x,t) = \wedge_{j=1}^J \bar{V}^{(j)}(x,t)$ satisfies

$$\bar{V}(x,T) \leq \infty \text{ for } x \in C^c \text{ and } \bar{V}(x,T) \leq 0 \text{ for } x \in C.$$

Example 1 Consider again the iid random example of Lecture 8, where $H(\alpha) = \log E e^{\alpha v_i}$ and $\{v_i, i \in \mathbb{N}\}$ are iid and mean zero. The set C in the example was of the form $(-\infty, \bar{\beta}] \cup [\beta^*, \infty)$, with $\bar{\beta} < 0 < \beta^*$. For this example a piecewise classical subsolution as the minimum of two functions is natural. One can easily construct solutions to the PDE of the simple form

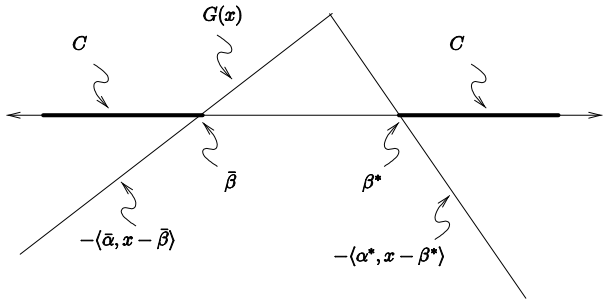

Figure 13: Terminal condition corresponding to a subsolution

$-ax + bt + c$ by requiring the relation $b + \mathbb{H}(-a) = b - H(a) = 0$. If $\hat{\alpha}$ and $\hat{\beta}$ are convex dual points, i.e.,

$$L(\hat{\beta}) = \sup_{\alpha \in \mathbb{R}} \Big[\alpha\hat{\beta} - H(\alpha)\Big] = \hat{\alpha}\hat{\beta} - H(\hat{\alpha})$$

we obtain the solution $-\hat{\alpha}(x - \hat{\beta}) + (L(\hat{\beta}) - \hat{\alpha}\hat{\beta})[1 - t]$, which correponds to the terminal condition $-\hat{\alpha}(x - \hat{\beta})$. Thus the two solutions

$$\begin{aligned} \bar{V}^{(1)}(x,t) &= -\alpha^*(x - \beta^*) + (L(\beta^*) - \alpha^*\beta^*)[1 - t], \\ \bar{V}^{(2)}(x,t) &= -\bar{\alpha}(x - \bar{\beta}) + (L(\bar{\beta}) - \bar{\alpha}\bar{\beta})[1 - t], \end{aligned}$$

which correspond to the terminal conditions indicated in Figure 13, generate a piecewise subsolution. Note that since they are convex dual points, α^* and $\bar{\alpha}$ generate changes of measure with the means β^* and $\bar{\beta}$, respectively. See Figure 14. Note also that the *subsolution* $\bar{V}(x,t)$ has a much simpler

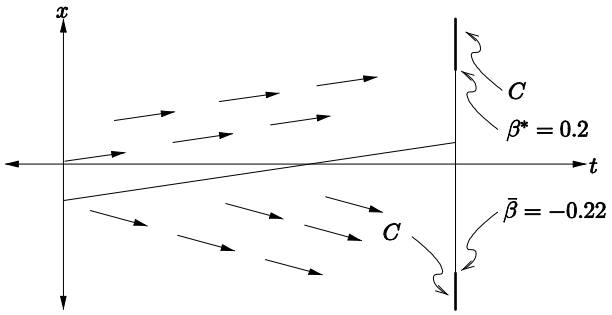

Figure 14: Partition of the domain by a piecewise classical subsolution

structure than the *solution* $V(x,t)$, but has the same value at $(0,0)$.

The definitions for the problem of entering a rare set B before a typical set A are similar. We consider $V : \mathbb{R}^d \to \mathbb{R}$.

$$\mathbb{H}(x, DV(x)) = 0, \tag{26}$$

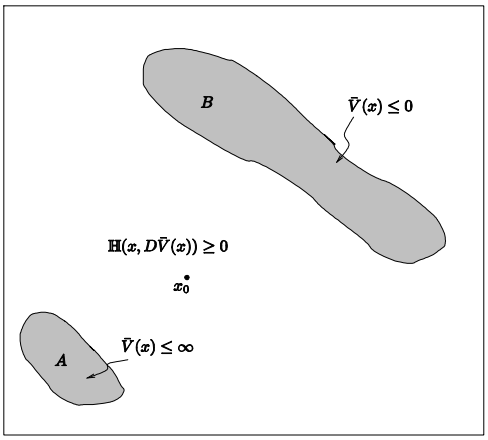

Figure 15: Subsolution for the exit problem

and the boundary conditions are

$$V(x) = \infty \text{ for } x \in \partial A \text{ and } V(x) = 0 \text{ for } x \in \partial B. \tag{27}$$

The importance sampling problem is to estimate

$$P_{x_0}\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

Definition 6 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a classical sense subsolution (or just classical subsolution) if it is continuously differentiable and if

$$\mathbb{H}(x, D\bar{V}(x)) \geq 0$$

for all $x \in (A \cup B)^c$, and if

$$\bar{V}(x) \leq \infty \text{ for } x \in A \text{ and } \bar{V}(x) \leq 0 \text{ for } x \in B.$$

Definition 7 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a piecewise classical sense subsolution (or just piecewise classical subsolution) if the following hold. For some $J \in \mathbb{N}$ there are functions $\bar{V}^{(j)} : \mathbb{R}^d \to \mathbb{R}, j = 1, \ldots, J$, that are continuously differentiable and satisfy

$$\mathbb{H}(x, D\bar{V}^{(j)}(x)) \geq 0$$

for all $x \in (A \cup B)^c$. Moreover $\bar{V}(x) = \wedge_{j=1}^{J} \bar{V}^{(j)}(x)$ satisfies

$$\bar{V}(x) \leq \infty \text{ for } x \in A \text{ and } \bar{V}(x) \leq 0 \text{ for } x \in B.$$

Of couse there are many other types of events and (risk-sensitive) expected values that one could consider, and the interested reader can find the appropriate definitions for many of these in the references. These two will suffice to illustrate the main points.

1.1 The IS scheme associated to a subsolution

Consider the finite time problem. As discussed in Lecture 8, if a smooth solution $V(x,t)$ to the HJB equation were available, then the correct change of measure if the current state of the simulated trajectory is at $\bar{X}_i^n$ would be to replace the original distribution on the noise $v_i(\bar{X}_i^n)$, i.e., $\theta(dy|\bar{X}_i^n)$, by

$$\nu_\alpha(dv|\bar{X}_i^n) = e^{\langle \alpha, v\rangle - H(\bar{X}_i^n,\alpha)}\theta(dv|\bar{X}_i^n) \text{ with } \alpha = -DV(\bar{X}_i^n, i/n).$$

If one is using a classical subsolution to design a scheme we follow exactly the same recipe, and the resulting second moment, rewritten in terms of the original random variables and process model, will equal

$$M(\bar{V}) \doteq E_{x_0}\left[1_{\{X_{Tn}^n \in C\}} \prod_{i=0}^{Tn-1} e^{\left\langle D\bar{V}(X_i^n,i/n), v_i(X_i^n)\right\rangle + H(X_i^n, -D\bar{V}(X_i^n,i/n))}\right].$$

If dealing with a piecewise classical sense subsolution, the situation is different. In such a case the gradient $D\bar{V}$ is not smooth, and the analysis used below to prove rigorous bounds on the performance would not apply. In this case we can mollify $\bar{V}$ and use a very simple-to-implement mixture of the changes of measure associated with the functions $\bar{V}^{(j)}$ appearing in $\bar{V}(x,t) = \wedge_{j=1}^J \bar{V}^{(j)}(x,t)$.

To be precise, for a small parameter $\delta > 0$ the standard mollification

$$\bar{V}^\delta(x,t) = -\delta \log\left(e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}\right)$$

is used. The properties of this mollification are summarized in the following lemma.

Lemma 1 *Let*

$$\bar{V}^\delta(x,t) = -\delta \log\left(e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}\right)$$

where each function $\bar{V}^{(j)}(x,t), j = 1,\ldots,J$ is continuously differentiable. Define the weights

$$\rho_j^\delta(x,t) = \frac{e^{-\frac{1}{\delta}\bar{V}^{(j)}(x,t)}}{e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}}.$$

Then

$$D\bar{V}^\delta(x,t) = \sum_{j=1}^J \rho_j^\delta(x,t) D\bar{V}^{(j)}(x,t) \text{ and } \bar{V}_t^\delta(x,t) = \sum_{j=1}^J \rho_j^\delta(x,t)\bar{V}_t^{(j)}(x,t)$$

Moreover

$$e^{-\frac{1}{\delta}\bar{V}(x,t)} \le e^{-\frac{1}{\delta}\bar{V}^\delta(x,t)} \le Je^{-\frac{1}{\delta}\bar{V}(x,t)},$$

and therefore

$$\bar{V}(x,t) \ge \bar{V}^\delta(x,t) \ge \bar{V}(x,t) - \delta \log J.$$

The role of the mollification is to define a mixture whose performance is very close to that of a classical subsolution, without giving up the flexibility and convenience of piecewise subsolutions. Specifically, it is implemented as follows. Given that the state of the current simulated trajectory is $\bar{X}_i^n$, we generate an independent random variable $\kappa_i^n \in \{1, \ldots, J\}$ with probabilities $\rho_j^\delta(\bar{X}_i^n, i/n)$, and then use the change of measure defined by $\alpha = -D\bar{V}^{(j)}(\bar{X}_i^n, i/n)$ if $\kappa_i^n = j$. The resulting second moment, rewritten in terms of the original process and noises, is then

$$M(\bar{V}) \doteq$$
$$E_{x_0}\left[1_{\{X_{Tn}^n \in C\}} \prod_{i=0}^{Tn-1} \left(\sum_{j=1}^{J} \rho_j^\delta \left(X_i^n, \frac{i}{n} \right) e^{\left\langle D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right), v_i(X_i^n)\right\rangle + H\left(X_i^n, -D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right)\right)} \right) \right].$$

The implementation and resulting form of the second moment are entirely analogous for the problem of hitting a rare set before a typical set, save that the scheme has no explicit dependence on time, and Tn is replaced by the first exit time N^n.

2 Statement of resulting performance

We recall that the performance of any scheme is characterized by the variance of a single sample, and that since the schemes are unbiased minimizing the variance is equivalent to minimizing the second moment. We also recall that the best possible rate of decay for this second moment is precisely twice the large deviation rate for the quantity being estimated.

We next identify the decay rate for a scheme constructed in terms of a subsolution, as described in the last section. As we will see, the rate has a very simple expression, and moreover the proof of this fact will follow almost immediately from the same argument used to prove the large deviation upper bound. We first state the result for the finite time problem and then the corresponding result for the exit problem. The process model will be the state dependent random walk model of Lecture 6. However, for the reason just given the proof carries over to other process models once one has established the corresponding large deviation theory

For the finite time problem we estimate $P_{x_0}\{X^n(T) \in C\}$, and to simplify the discussion we will want a large deviation limit to hold. This requires some regularity of C. For example, such a limit will hold if C is the closure of its interior.

Theorem 2 *Consider the process model $\{X^n, n \in \mathbb{N}\}$ of Lecture 6 and assume*

$$\begin{aligned} -\frac{1}{n}\log P_{x_0}\{X^n(T) \in C\} &\to V(0, x_0) \\ &= \inf\{I_T(\phi) : \phi(0) = x_0, \phi(T) \in C\}. \end{aligned}$$

Let $\bar{V}$ be a classical subsolution for the corresponding PDE and define the IS scheme as in the last section. Then the second moment for this scheme satisfies

$$\liminf -\frac{1}{n}\log M(\bar{V}) \geq V(x_0,0) + \bar{V}(x_0,0).$$

If $\bar{V}$ is a piecewise classical subsolution and if $\bar{V}^\delta$ is the corresponding mollification, then the last display holds with $\bar{V}$ replaced by $\bar{V}^\delta$.

Remark 1

1. The conclusion of the result is clear. The performance of the scheme based on any subsolution is measured by the value of the subsolution at the starting point, with larger values giving better performance. When there is a comparison principle for the PDE a subsolution can never be greater than the solution, and the best possible value is $\bar{V}(x_0,0) = V(x_0,0)$, which corresponds to asymptotic optimality.

2. Note that the equality between subsolution and solution is only required at the starting point $(x_0,0)$. For many problems subsolutions with the optimal value at one point are structrually simpler and much easier to find than the solution. Also, there can be many subsolutions with the optimal value at the starting point.

3. Note that the subsolution $\bar{V}(x_0,0) \equiv 0$ corresponds to standard Monte Carlo, and gives the very poor rate of decay $V(x_0,0)$. Thus any subsolution with $\bar{V}(x_0,0) > 0$ will improve on standard Monte Carlo, though it is also possible that a scheme could correspond to $\bar{V}(x_0,0) < 0$ and do even worse than standard Monte Carlo!

4. For the piecewise classical subsolution, one can allow $\delta = \delta_n \to 0$ as $n \to \infty$ to get a limit in terms of $\bar{V}$ rather than $\bar{V}^\delta$, so long as $n\delta_n \to \infty$.

5. In the case where the piecewise subsolution is constructed from exactly two pieces one can show that the mollification is not needed.

3 Example

A subsolution with the optimal value at the origin was identified in Example 1 for the problem used to illustrate the shortcomings of the "open loop" sampling schemes discussed in Lecture 8. The sampling scheme based on the piecewise subsolution incorporates state feedback, and in fact will change the distribution of the next increment as illustrated in Figure 14. The theoretical results of Theorem 2 are reflected in the performance of the scheme–see Table 1. We recall that the true value for the problam data is $p_n = 3.67 \times 10^{-2}$.

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}_n$ ($\times 10^{-2}$)	3.72	3.65	3.67	3.56
Standard Error ($\times 10^{-2}$)	0.11	0.11	0.10	0.10
95% Confidence Interval ($\times 10^{-2}$)	[3.50, 3.94]	[3.43, 3.87]	[3.47, 3.87]	[3.36, 3.76]

Figure 16: Table 1

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}_n$ ($\times 10^{-2}$)	2.23	2.24	17.32	16.37
Standard Error ($\times 10^{-2}$)	0.05	0.05	14.98	14.10
95% C.I. ($\times 10^{-2}$)	[2.13, 2.33]	[2.14, 2.34]	[-12.64, 47.28]	[-11.83, 44.57]

Figure 17: Table 2

This should be contrasted with the performance of the open loop sampling scheme presented in Table 2.

Lecture 10: The Subsolutions Approach to Importance Sampling, Cont'd

1 Proof of the performance bound for the finite time problem

We recall the statement of the theorem regarding performance of a IS scheme based on a subsolution $\bar{V}$.

Theorem 1 *Consider the process model $\{X^n, n \in \mathbb{N}\}$ of Lecture 6 and assume*

$$\begin{aligned} -\frac{1}{n}\log P_{x_0}\left\{X^n(T) \in C\right\} &\rightarrow V(0,x_0) \\ &= \inf\left\{I_T(\phi) : \phi(0) = x_0, \phi(T) \in C\right\}. \end{aligned}$$

Let $\bar{V}$ be a classical subsolution for the corresponding PDE and define the IS scheme as in the last section. Then the second moment for this scheme satisfies

$$\liminf -\frac{1}{n}\log M(\bar{V}) \geq V(x_0,0) + \bar{V}(x_0,0).$$

If $\bar{V}$ is a piecewise classical subsolution and if $\bar{V}^\delta$ is the corresponding mollification, then the last display holds with $\bar{V}$ replaced by $\bar{V}^\delta$.

Proof. We first consider the simpler case of a classical subsolution. Recall from Lecture 8 that when rewritten in terms of the original process model, the second moment of the scheme defined in terms of a subsolution $\bar{V}$ took the form

$$M(\bar{V}) = E_{x_0}\left[1_{\left\{X^n_{Tn}\in C\right\}} \prod_{i=0}^{Tn-1} e^{\left\langle D\bar{V}(X_i^n,i/n), v_i(X_i^n)\right\rangle + H(X_i^n, -D\bar{V}(X_i^n,i/n))}\right].$$

As done many times before, we will rewrite this using a relative entropy representation. Thus we have

$$\begin{aligned} &-\frac{1}{n}\log M(\bar{V}) \\ &= \frac{1}{n}\inf_{\{\bar{\mu}_i^n\}} E_{x_0}\left[\sum_{i=0}^{Tn-1}\left[-\left\langle D\bar{V}\left(\bar{X}_i^n, \frac{i}{n}\right), \bar{v}_i\right\rangle - H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, \frac{i}{n}\right)\right)\right]\right. \\ &\qquad \left. + \sum_{i=0}^{Tn-1} R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\,\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\left\{\bar{X}^n_{Tn}\in C^c\right\}}\right]. \end{aligned}$$

Our goal is to prove a lower bound for this quantity. In fact, virtually all that is needed has already been established in proving the large deviation

upper bound for the sequence $\{X^n, n \in \mathbb{N}\}$. First note that since $\bar{V}$ is a subsolution,

$$-H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) = \mathbb{H}\left(\bar{X}_i^n, D\bar{V}\left(\bar{X}_i^n, i/n\right)\right) \geq -\bar{V}_t(\bar{X}_i^n, i/n). \quad (28)$$

Next note that we can assume without loss that the expected relative entropy is bounded from above, since otherwise there is nothing to prove. We can therefore use the results on tightness of controls and controlled processes as well as their asymptotic relations proved in Lecture 6. To be specific, we use recall the controlled empirical measures

$$\bar{L}^n(A \times B) \doteq \int_B \bar{L}^n(A|t)dt, \quad \bar{L}^n(A|t) = \delta_{\bar{v}_i^n}(A) \text{ if } t \in [i/n - 1/n, in),$$

as well as the piecewise constant interpolation

$$\hat{X}^n(t) = \bar{X}_i^n, \quad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right).$$

It was proved in Lecture 6 that $\left\{(\bar{X}^n, \bar{L}^n, \bar{\mu}^n), n \in \mathbb{N}\right\}$ was tight, where $\bar{X}^n$ is the piecewise linear interpolation, $\bar{\mu}^n$ is the control measure, and that also y is uniformly integrable with respect to $\bar{L}^n(dy \times dt)$. It was also shown that any limit $(\bar{X}, \bar{L}, \bar{\mu})$ allowed the decomposition $\bar{\mu}(dy|t)dt$, and that

$$\bar{X}(t) = \int_{\mathbb{R}^d \times [0,t]} y\bar{L}(dy \times dt) + x_0 = \int_{\mathbb{R}^d \times [0,t]} y\bar{\mu}(dy|t)dt + x_0$$

It is easy to check that the piecewise linear and piecewise constant interpolations $\bar{X}^n$ and $\hat{X}^n$ must have the same limit.

We can now argue why the lower bound follows from these previously proved results. Consider any weakly converging subsequence, which we again denote by n. Exactly as in Lecture 6 we have

$$\begin{aligned} &\liminf_{n\to\infty} E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{Tn-1} R\left(\bar{\mu}_i^n(\cdot)\,\|\theta(\cdot|\bar{X}_i^n)\right) + \infty 1_{\left\{\bar{X}_{Tn}^n \in C^c\right\}}\right] \\ &\quad \geq \quad E_{x_0}\left[\int_0^T L(\bar{X}, \dot{\bar{X}})dt + \infty 1_{\left\{\bar{X}(T)\in C^c\right\}}\right]. \end{aligned}$$

We have already explained why the $-H\left(\bar{X}_i^n, -D\bar{V}\left(\bar{X}_i^n, i/n\right)\right)$ term may be replaced by $-\bar{V}_t(\bar{X}_i^n, i/n)$. To bring in the empirical measure on the $\bar{v}_i$, we rewrite the sum as

$$\begin{aligned} &E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{Tn-1}\left(\left\langle D\bar{V}\left(\bar{X}_i^n, \frac{i}{n}\right), \bar{v}_i\right\rangle - \bar{V}_t\left(\bar{X}_i^n, \frac{i}{n}\right)\right)\right] \\ &= \quad E_{x_0}\left[\int_{\mathbb{R}^d\times[0,T]}\left(\left\langle D\bar{V}\left(\hat{X}^n(t), t\right), y\right\rangle - \bar{V}_t\left(\hat{X}^n(t), t\right)\right)\bar{L}^n(dy \times dt)\right] + O(1/n), \end{aligned}$$

where the error term $O(1/n)$ is due to replacing $D\bar{V}\left(X_i^n, 1/n\right)$ by $D\bar{V}(\hat{X}^n(t), t)$ for $t \in [i/n, i/n + 1/n)$. Using the uniform integrability thus gives

$$\begin{aligned}
&\lim_{n\to\infty} E_{x_0}\left[\frac{1}{n}\sum_{i=0}^{Tn-1}\left(\left\langle D\bar{V}\left(\bar{X}_i^n, \frac{i}{n}\right), \bar{v}_i\right\rangle - \bar{V}_t\left(\bar{X}_i^n, \frac{i}{n}\right)\right)\right] \\
&= E_{x_0}\left[\int_{\mathbb{R}^d\times[0,T]}\left(\left\langle D\bar{V}\left(\bar{X}, t\right), y\right\rangle - \bar{V}_t\left(\bar{X}, t\right)\right)\bar{\mu}(dy|t)dt\right] \\
&= E_{x_0}\left[\int_0^T\left\langle D\bar{V}\left(\bar{X}, t\right), \dot{\bar{X}}\right\rangle dt\right].
\end{aligned}$$

Hence we have the combined lower bound

$$E_{x_0}\left[\int_0^T\left(-\left\langle D\bar{V}\left(\bar{X}, t\right), \dot{\bar{X}}\right\rangle - \bar{V}_t\left(\bar{X}, t\right)\right)dt + \int_0^T L(\bar{X}, \dot{\bar{X}})dt + \infty 1_{\left\{\bar{X}(T)\in C^c\right\}}\right].$$

Unless $\bar{X}(T) \in C$ w.p.1 this cost is infinite. Using that $\bar{X}(0) = x_0$ and that since $\bar{V}$ is a subsolution $\bar{V}(x, T) \leq 0$ for $x \in C$, by the (ordinary) chain rule

$$\int_0^T\left(-\left\langle D\bar{V}\left(\bar{X}, t\right), \dot{\bar{X}}\right\rangle - \bar{V}_t\left(\bar{X}, t\right)\right)dt \geq \bar{V}(x_0, 0)$$

and from the definition of $V(x_0, 0)$

$$\int_0^T L(\bar{X}, \dot{\bar{X}})dt \geq V(x_0, 0),$$

both w.p.1. We thus obtain the lower bound $V(x_0, 0) + \bar{V}(x_0, 0)$, which concludes the proof for the case of a classical subsolution.

The proof for the piecewise classical is very similar. In this case the second moment is

$$E_{x_0}\left[1_{\left\{X_{Tn}^n\in C\right\}}\prod_{i=0}^{Tn-1}\left(\sum_{j=1}^J \rho_j^\delta\left(X_i^n, \frac{i}{n}\right)e^{\left\langle D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right), v_i(X_i^n)\right\rangle + H\left(X_i^n, -D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right)\right)}\right)\right].$$

Using Jensen's inequality gives the lower bound

$$E_{x_0}\left[1_{\left\{X_{Tn}^n\in C\right\}}\prod_{i=0}^{Tn-1}\left(e^{\sum_{j=1}^J \rho_j^\delta\left(X_i^n, \frac{i}{n}\right)\left[\left\langle D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right), v_i(X_i^n)\right\rangle + H\left(X_i^n, -D\bar{V}^{(j)}\left(X_i^n, \frac{i}{n}\right)\right)\right]}\right)\right].$$

For this we write a representation in the usual way. The relative entropy terms in this representation are treated in exactly the same way as was just done for a classical subsolution. We also use that $\bar{V}^{(j)}$ satisfies (28) to replace $-H\left(\bar{X}_i^n, -D\bar{V}^{(j)}\left(\bar{X}_i^n, i/n\right)\right)$ by the smaller quantity $-\bar{V}_t^{(j)}(\bar{X}_i^n, i/n)$. Then the new and different terms are of the form

$$-\int_0^T\sum_{j=1}^J \rho_j^\delta\left(\bar{X}, t\right)\left(\left\langle D\bar{V}^{(j)}\left(\bar{X}, t\right), \dot{\bar{X}}\right\rangle + \bar{V}_t^{(j)}\left(\bar{X}, t\right)\right)dt.$$

However, using the identities

$$D\bar{V}^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) D\bar{V}^{(j)}(x,t) \text{ and } \bar{V}_t^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) \bar{V}_t^{(j)}(x,t)$$

from Lemma 1 of Lecture 9, we can apply the chain rule as before and with exactly the same result. □

Remark 1 The argument uses little of the particular properties of the underlying process, given that one has used the weak convergence to establish the large deviation upper bound, and thus adapts with few changes to other process models.

Remark 2 The proof for the problem of entering a rare set B before a typical set A is essentially reduced to the finite time problem. Recall that it is assumed that there is a single global attractor that is contained in the interior of A. As in [34, Lemma 2.2, Chapter 5], one can argue that there is $K(T) \to \infty$ as $T \to \infty$ such that the large deviation rate for any trajectory that starts at x_0 and enters neither A nor B by time T has cost at least $K(T)$. Using this one can argue there is an arbitrarily large lower bound in the representation for the second moment for trajectories that do not reach A or B by time T. Lower bounds for the remaining trajectories follow as in the finite time case from the chain rule for $\bar{V}$ and the definition of V.

2 Examples

2.1 Level crossing

We consider Example 1 of Lecture 1 with $Z_i \sim (\text{Exp}(1) - 2, \text{Exp}(1) - 3)$, where Exp(1) denotes the exponential distribution with mean 1 (this allows the exact solution to be explicitly computed). The subsolution (actually a solution) is constructed as in Example 4 of Lecture 7. The first table gives the outcome from the state independent scheme, and the second table presents 4 sets of 20,000 simulations each, where the true value is 9.51×10^{-5}. The third and fourth tables give the corresponding results for the scheme based on the subsolution, with mollification parameters $\delta = 0.1$ and 0.2, respectively. The example is taken from [29].

2.2 Server slowdown

Here we consider Example 4 of Lecture 1, with parameters $(\lambda, \mu_1, \mu_2, \nu_1) = (0.3, 0.36, 0.34, 0.32)$. The construction of a subsolution suitable for importance sampling is more complex than the last example, since it includes

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	9.51×10^{-5}	2.88×10^{-8}	9.24×10^{-12}
Estimate	7.13×10^{-5}	2.44×10^{-8}	8.36×10^{-12}
Standard Error	0.06×10^{-5}	0.02×10^{-8}	0.08×10^{-12}
95% C.I.	$[7.01, 7.25] \times 10^{-5}$	$[2.40, 2.48] \times 10^{-8}$	$[8.20, 8.52] \times 10^{-12}$

	No. 1	No. 2	No. 3	No. 4
Estimate ($\times 10^{-5}$)	7.12	23.4	53.9	7.01
Standard Error ($\times 10^{-5}$)	0.07	16.3	46.9	0.07
95% C.I. ($\times 10^{-5}$)	[6.98,7.26]	[-9.2,56.0]	[-39.9, 147.7]	[6.87, 7.15]

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	9.51×10^{-5}	2.88×10^{-8}	9.24×10^{-12}
Estimate	9.56×10^{-5}	2.87×10^{-8}	9.31×10^{-12}
Standard Error	0.10×10^{-5}	0.03×10^{-8}	0.09×10^{-12}
95% C.I.	$[9.36, 9.76] \times 10^{-5}$	$[2.81, 2.93] \times 10^{-8}$	$[9.13, 9.49] \times 10^{-12}$

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	9.51×10^{-5}	2.88×10^{-8}	9.24×10^{-12}
Estimate	9.54×10^{-5}	2.90×10^{-8}	9.14×10^{-12}
Standard Error	0.10×10^{-5}	0.03×10^{-8}	0.11×10^{-12}
95% C.I.	$[9.34, 9.74] \times 10^{-5}$	$[2.84, 2.96] \times 10^{-8}$	$[8.92, 9.36] \times 10^{-12}$

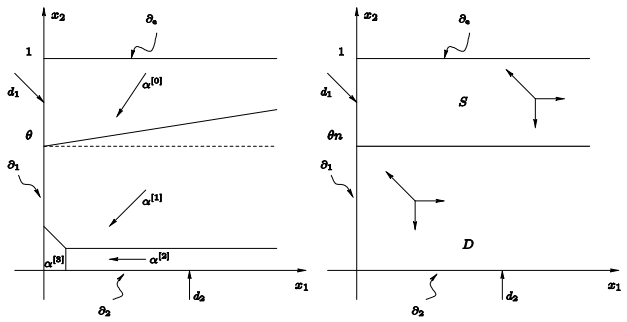

Figure 18: Subsolution based on 4 affine functions

regions of differing statistical behavior as well as boundaries, see [21]. The construction is based on the roots for various Hamiltonians, and leads to a subsolution that is the minimum of 4 affine functions, whose gradients are depicted as $\alpha^{[j]}$,$j = 0, 1, 2, 3$ in the left side Figure 1. The true solution is much more complex. The resulting performance is given in the table, based on $K = 1,000,000$ samples.

	$n = 20$	$n = 50$	$n = 100$
Theoretical value	5.63×10^{-2}	1.19×10^{-3}	1.63×10^{-6}
Estimate	5.62×10^{-2}	1.18×10^{-3}	1.61×10^{-6}
Std. Err.	0.03×10^{-2}	0.01×10^{-3}	0.02×10^{-6}
95% C.I.	$[5.56, 5.68] \times 10^{-2}$	$[1.16, 1.20] \times 10^{-3}$	$[1.57, 1.65] \times 10^{-6}$

2.3 Path dependent functional

Let $\{Y_1, Y_2, \ldots\}$ be a sequence of iid random variables with common distribution μ and $E[Y_i] = 0$. As before, let H be the log-moment generating function and L its convex conjugate. Fix $n \in \mathbb{N}$, and for $1 \leq i \leq n$ define

$$X_i^n \doteq \frac{1}{n} \sum_{j=1}^{i} Y_j,$$

with $X_0^n \doteq 0$. We are interested in estimating

$$E_n \doteq E\left[e^{-nF(X_n^n)} 1_{\{\max_{0 \leq i \leq n} X_i^n \geq h\}}\right]$$

where $h > 0$ is a given constant. Assume that the large deviation limit

$$\lim_{n \to \infty} -\frac{1}{n} \log E_n = \gamma$$

holds, with γ the solution of the following variational problem:

$$\gamma = \inf \left\{ \int_0^1 L(\dot{\phi}(t))\, dt + F(\phi(1)) : \phi \in \mathcal{A}_0([0,1] : R), \max_{0 \le t \le 1} \phi(t) \ge h \right\}.$$

To write down the PDE associated with this estimation problem, one needs to expand the state space to accommodate the path-dependence of the event. More precisely, the state process is (X_i^n, B_i^n), where

$$B_i \doteq 1_{\{\max_{0 \le j \le i} X_j^n \ge h\}}$$

is the indicator of whether or not the "barrier" h has been breached by time i. One obtains a coupled pair of PDEs for the subsolution, of the form

$$\bar{V}_t(1,x,t) + \mathbb{H}(D\bar{V}(1,x,t)) \ge 0, \quad \bar{V}(1,x,t) \le 2F(x),$$

$$\bar{V}_t(0,x,t) + \mathbb{H}(D\bar{V}(0,x,t)) \ge 0, \quad \bar{V}(0,x,t) \le \bar{V}(1,x,t)$$

for $x \ge h, t \in [0,1]$.

We consider the specific problem $E_n \doteq E[1_{\{\max_{0 \le i \le n} X_i^n \ge h\}} 1_{\{X_n^n \le l\}}]$ with $l < h$. Again a subsolution with the optimal value at the origin can be constructed, for details see [29]. For the numerical example we take $Y_i \sim N(0,1)$ and $h = 1$ and $l = 0.8$. Simulations were run for $n = 10, 20, 30$, and each estimate consists of 20,000 samples. What we call the "theoretical value" is an estimate based on 1 billion samples of the importance sampling scheme.

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.68×10^{-5}	9.66×10^{-9}	6.09×10^{-12}
Estimate	1.74×10^{-5}	9.58×10^{-9}	6.26×10^{-12}
Standard Error	0.04×10^{-5}	0.27×10^{-9}	0.19×10^{-12}
95% C.I.	$[1.66, 1.82] \times 10^{-5}$	$[9.04, 10.12] \times 10^{-9}$	$[5.88, 6.64] \times 10^{-12}$

2.4 Other examples

Subsolutions have been constructed for many other types of problems, including the following.

- Networks with feedback
- Non-Markovian systems
- Serve-the-longer discipline
- Open/closed network

- General Jackson networks
- Reversible systems
- Multi-scale processes (homogenization)

For details, see [21, 23, 25, 27, 29, 30].

Lecture 11: The Empirical Measure of a Markov Chain

1 Problem formulation

In this lecture we consider the large deviation theory for the empirical measure of a Markov chain, thus generalizing Sanov's Theorem from Lecture 3. The ideas developed here are useful in other contexts, such as proving sample path large deviation properties of processes whose "driving noises" have a Markovian structure rather than an iid structure.

To focus on the main issues, we simplify by considering only Markov chains with a compact state space S. Dealing with the unbounded case requires existence of a suitable Lyapunov function to proves the required tightness results [14, 12, 13]. These results are automatic when the state space is compact.

Thus let $\{X_i, i \in \mathbb{N}_0\}$ denote a Markov chain with transition kernel $p(x, dy)$ and compact state space S. The object of interest is the empirical measure defined by

$$L^n(A) = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{X_i}(A) \text{ for } A \in \mathcal{B}(S).$$

Under ergodicity there will be a unique invariant measure $\pi \in \mathcal{P}(S)$, and by the ergodic theorem we have the LLN result $L^n \to \pi$ in the weak topology, w.p.1. The goal is to prove an LDP for $\{L^n, n \in \mathbb{N}\}$ in the weak topology, and identify the rate function.

Remark 1 Although both here and when considering Sanov's Theorem we use the weak topology, the results also hold in the stronger τ-topology [14, Section 9.3]. As a consequence, one can approximate the distribution of quantities such as $\int_S f(x) L^n(dx)$ when f is just bounded and measurable (and not necessarily continuous).

2 Some applications

2.1 Markov modulated dynamics

Many of the process models mentioned in Lecture 1 can be made more realistic for applications by allowing the distribution of the driving noise to depend on an exogenous Markov chain. For example, Example 1 models insurance risk, with the noises Z_i representing the difference between income and payouts at time i. A more realistic model would allow the distribution μ to depend on a finite state Markov chain Y_i representing, e.g., the state of the economy and other factors. Such a process would be called *Markov modulated*. Similarly, in order to model "bursty" data the arrival rate in the

queueing and data loss problem (Example 4) should depend on a finite state chain. For such more sophisticated process models the rate function can be found and a large deviation analysis given by combining the methods used in say Lecture 6 with those of the present section. The construction of IS schemes for these processes is also possible, and various examples can be found in the references at the end of Lecture 10.

2.2 Markov chain Monte Carlo

One of the most important uses of the empirical measure is in the numerical approximation of integrals of the form $\int_S f(x)\pi(dx)$, and in particular when π is a Gibbs measure, i.e., a measure of the form $e^{-V(x)/\tau}dx/Z$, where V is a potential function, τ is a parameter, and Z is a normalization that makes the indicated measure a probability measure. There are well known methods to construct ergodic Markov processes $\{X_i, i \in \mathbb{N}_0\}$ for which π is the unique invariant distribution, and thus $\int_S f(x)L^n(dx)$ gives a converging approximation to $\int_S f(x)\pi(dx)$. This technique has a tremendous number of very practical applications in the physical and biological sciences, engineering, statistics, and elsewhere.

However, for many problems the dimension of S is very large, and in addition the methods that generate the chain from V have the property that when V has many deep local minima, parts of the state space communicate poorly under the dynamics $p(x, dy)$. When this happens, and it happens very frequently, the problem of algorithm design becomes crucial.

In order to compare algorithms, one needs a criterion for good performance. Since it focuses on the object of interest, i.e., the empirical measure, it would seem that the large deviation rate is a good measure. The rate function I depends of course on the dynamics, though for any chain leading to π as an invariant distribution $I(\mu) = 0$ if and only $\mu = \pi$. Different algorithms lead to different rate functions, the rate functions give one a great deal of information that can be used to compare the algorithms.

This should be compared with other measures that have been traditionally used to compare chains, such as the second eigenvalue. Let $p(x, dy)$ be an ergodic transition kernel with invariant distribution π. Then $p(x, \cdot)$ has a single eigenvalue of modulus 1 corresponding to the eigenvector π, and the magnitude $|\lambda_1|$ of the next largest is often used to characterize the performance of the associated empirical measure. However, the second eigenvalue provides information only on convergence of the n-step transition kernel $p^{(n)}(x, dy) = P\{X_n \in dy | X_0 = x\}$, and does not give any direct information on the empirical measure.

A recent work which effectively applied the large deviation rate as a measure of rate of convergence is [24], and its further application to problems of algorithm design are ongoing.

3 The representation

To apply the weak convergence approach, we first need the representation. As usual, it will follow from the high level representation (Lemma 4 of Lecture 2) and the chain rule (Lemma 5). The base measure in this case is the Markov measure

$$\bar{p}^n(x_0, dx_1, \ldots, dx_n) = p(x_0, dx_1)p(x_1, dx_2)\cdots p(x_{n-1}, dx_n)$$

on S^n. We consider any measure $\mu^n(x_0, dx_1, \ldots, dx_n)$ with finite relative entropy with respect to $\bar{p}^n$, and factor it as the conditional distribution on the jth variable given all preceding variables:

$$\begin{aligned}&\mu^n(x_0, dx_1, \ldots, dx_n)\\ &\quad = \mu^n_{1|0}(dx_1|x_0)\mu^n_{2|0,1}(dx_2|x_0, x_1)\cdots\mu^n_{n|0,\ldots,n-1}(dx_n|x_0, \ldots, x_{n-1}).\end{aligned}$$

Let $(\bar{X}^n_1, \ldots, \bar{X}^n_n)$ have the distribution μ^n. As was the case with Sanov's Theorem, we define the random control measures

$$\bar{\mu}^n_i(dx_i) = \mu^n_{i|0,\ldots,i-1}(dx_i|x_0, \bar{X}^n_1 \ldots, \bar{X}^n_{i-1}),$$

so that $\bar{\mu}^n_i(dx_i)$ picks the distribution of $\bar{X}^n_i$ given $\bar{X}^n_1 \ldots, \bar{X}^n_{i-1}$. Finally, let $\bar{L}^n$ be the controlled empirical measure (with $\bar{X}^n_0 = x_0$):

$$\bar{L}^n(A) = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{\bar{X}^n_i}(A) \text{ for } A \in \mathcal{B}(S).$$

Then using the chain rule as was done previously (see Lecture 2), we have

$$-\frac{1}{n}\log Ee^{-nF(L^n)} = \inf_{\{\bar{\mu}^n_i\}} E\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^{n} R\left(\bar{\mu}^n_i(\cdot)\,\middle\|\,p(\bar{X}^n_{i-1}, \cdot)\right)\right]$$

for any bounded and measurable $F : \mathcal{P}(S) \to \mathbb{R}$. To prove an LDP it will be enough to consider bounded and continuous F.

3.1 Form of the rate function

In the setting of Sanov's Theorem the minimizing controls were found, a posteriori, to be asymptotically product measure, reflecting the form of the base measure on the collection $\{X_i, i \in \mathbb{N}_0\}$. One might suspect something analogous here, which is that nearly optimizing controls for large n might be of the Markov form $\bar{\mu}^n_i(dx_i) = q(\bar{X}^n_{i-1}, dx_i)$ for some transition kernel q. With this in mind, we rewrite the relative entropy using the chain rule:

$$\begin{aligned}R\left(\bar{\mu}^n_i(\cdot)\,\middle\|\,p(\bar{X}^n_{i-1}, \cdot)\right) &= R\left(\bar{\mu}^n_i(\cdot)\,\middle\|\,p(\bar{X}^n_{i-1}, \cdot)\right) + R\left(\delta_{\bar{X}^n_{i-1}}(\cdot)\,\middle\|\,\delta_{\bar{X}^n_{i-1}}(\cdot)\right)\\ &= R\left(\delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy)\,\middle\|\,\delta_{\bar{X}^n_{i-1}}(dx)p(x, dy)\right).\end{aligned}$$

The measure $\delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy)$ records the control used to pick the distribution of $\bar{X}^n_i$ depending on the location of $\bar{X}^n_{i-1}$, and will help us learn the form of $q(x, dy)$.

Let us see if we can guess the form of the rate function, and at the same time give the proof of the lower bound (the large deviation upper bound). By Jensen's inequality and the joint convexity of relative entropy

$$\begin{aligned}
& E\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy) \,\middle\|\, \delta_{\bar{X}^n_{i-1}}(dx)p(x,dy)\right)\right] \\
&\geq \quad E\left[F(\bar{L}^n) + R\left(\frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy) \,\middle\|\, \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}^n_{i-1}}(dx)p(x,dy)\right)\right] \\
&= \quad E\left[F(\bar{L}^n) + R\left(\frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy) \,\middle\|\, \bar{L}^n(dx)p(x,dy)\right)\right].
\end{aligned}$$

Let

$$\lambda^n(dx \times dy) = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}^n_{i-1}}(dx)\bar{\mu}^n_i(dy).$$

Since S and hence S^2 are compact so are $\mathcal{P}(S)$ and $\mathcal{P}(S^2)$, and so automatically $\left\{\left(\lambda^n, \bar{L}^n\right), n \in \mathbb{N}\right\}$ is tight. Note that $\bar{L}^n$ is the first marginal of λ^n, and that since $\bar{\mu}^n_i(dy)$ is picking the distribution of $\bar{X}^n_i$, the martingale generalization of the Glivenko-Cantelli lemma (Lemma 4 in Lecture 3) shows that asymptotically the first and second marginals of λ^n are the same. Thus if $\left(\lambda^n, \bar{L}^n\right) \to \left(\lambda, \bar{L}\right)$ in distribution along a subsequence, then $[\lambda]_1(dx) = [\lambda]_2(dx) = \bar{L}(dx)$, where $[\lambda]_1$ and $[\lambda]_2$ denote the first and second marginals of λ. We will assume that $p(x, dy)$ is *Feller*, i.e., that the mapping $x \to p(x, \cdot)$ is continuous in the topology of weak convergence. This will imply $\bar{L}^n(dx)p(x, dy) \to \bar{L}(dx)p(x, dy)$ in distribution. Indeed, if $f(x, y)$ is bounded and continuous, then by the Feller property $x \to \int_S f(x, y)p(x, dy)$ is bounded and continuous, and since

$$\int_S \int_S f(x,y)p(x,dy)\bar{L}^n(dx) \to \int_S \int_S f(x,y)p(x,dy)\bar{L}(dx)$$

and f is arbitrary the result follows.

We can then compute a lower bound along the weakly converging subsequence using Fatou's Lemma, the continuity of F, and lower semicontinuity of R:

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} \\
&\quad \geq \quad \liminf_{n\to\infty} E\left[F(\bar{L}^n) + R\left(\lambda^n(dx \times dy) \,\middle\|\, \bar{L}^n(dx)p(x,dy)\right)\right] \\
&\quad \geq \quad E\left[F(\bar{L}) + R\left(\lambda(dx \times dy) \,\middle\|\, \bar{L}(dx)p(x,dy)\right)\right].
\end{aligned}$$

Suppose we define

$$I(\mu) = \inf_{\gamma \in \mathcal{P}(S^2), [\gamma]_1 = [\gamma]_2 = \mu} R\left(\gamma(dx \times dy) \| \mu(dx)p(x, dy)\right).$$

Then since $[\lambda]_1 = [\lambda]_2 = \bar{L}$ we have shown

$$\liminf_{n \to \infty} -\frac{1}{n} \log E e^{-nF(L^n)} \geq \inf_{\mu} \left[F(\mu) + I(\mu)\right], \tag{29}$$

suggesting that I may in fact be the rate function.

To complete the proof we must show the reverse inequality with the same function I. Note that if $[\gamma]_1 = [\gamma]_2 = \mu$, then we can factor $\gamma(dx \times dy)$ in the form $\gamma(dx \times dy) = \mu(dx)q(x, dy)$ for some transition kernel q, and that $[\gamma]_2 = \mu$ is exactly the statement that μ is an invariant distribution for q. This will suggest how to construct a control for the reverse inequality.

4 Assumptions and statement of the LDP

Condition 1 *S is compact and p satisfies the Feller property. In addition, p satisfies the following transitivity condition. There exist positive integers l_0 and n_0 such that for all x and ζ in S*

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i} p^{(i)}(x, dy) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} p^{(j)}(\zeta, dy), \tag{30}$$

where $p^{(k)}$ denotes the k-step transition probability.

Theorem 1 *Assume the condition just given on the Markov chain $\{X_i, i \in \mathbb{N}_0\}$ and let $\{L^n, n \in \mathbb{N}\}$ be the empirical measure. Then $\{L^n, n \in \mathbb{N}\}$ satisfies an LDP with rate function*

$$I(\mu) = \inf_{\gamma \in \mathcal{P}(S^2), [\gamma]_1 = [\gamma]_2 = \mu} R\left(\gamma(dx \times dy) \| \mu(dx)p(x, dy)\right).$$

Let $d(\cdot, \cdot)$ denote the metric on S. Suppose in addition that for each $x \in S$ there are $\beta \in \mathcal{P}(S)$, $k \in \mathbb{N}$, and $c > 0$ such that for all $\zeta \in S$ satisfying $d(\zeta, x) < c$ and all $A \in \mathcal{B}(S)$

$$\sum_{i=0}^{k} p^{(j)}(\zeta, A) \geq c\beta(A).$$

Then the large deviation estimates are uniform in the initial condition x_0.

5 Sketch of the proof

First note that under the compactness of S and the Feller condition $p(x, dy)$ has an invariant distribution π. Indeed, it is automatic that $\{EL^n(dx), n \in \mathbb{N}\}$ is tight and for any bounded and continuous f

$$\begin{aligned}
&\left| \int_S f(x) EL^n(dx) - \int_{S^2} f(y) p(x, dy) EL^n(dx) \right| \\
&= \left| E \int_S f(x) L^n(dx) - E \int_{S^2} f(y) p(x, dy) L^n(dx) \right| \\
&\leq \frac{2}{n} \|f\|_\infty \\
&\to 0.
\end{aligned}$$

If π is any weak limit then $\int_S f(x)\pi(dx) = \int_{S^2} f(y)p(x, dy)\pi(dx)$ follows from the Feller property, and thus π is invariant. Second note that the definition of I is as the infimum of a convex and lower semicontinuous function subject to an affine constraint. As a consequence, I is convex and lower semicontinuous. Since $\mathcal{P}(S)$ is compact, I has compact level sets.

The bound (29) has already been proved and all that remains is the upper bound

$$\limsup_{n\to\infty} -\frac{1}{n} \log E e^{-nF(L^n)} \leq \inf_\mu \left[F(\mu) + I(\mu) \right], \tag{31}$$

which is equivalent to the large deviation lower bound. We now state two key facts which follow from the transitivity condition whose proof will be omitted (see [14, Lemma 8.6.2]). The first is that $p(x, dy)$ is ergodic and π is unique, and the second is that $I(\mu) < \infty$ implies $\mu \ll \pi$. Note that from the definition $I(\pi) = 0$, as expected.

We now show how the bound (31) follows from these facts. Let $\varepsilon > 0$ and let μ^* satisfy $F(\mu^*)+I(\mu^*) \leq \inf_\mu [F(\mu) + I(\mu)]+\varepsilon$. From the definition of I there is $\gamma \in \mathcal{P}(S^2)$ such that $[\gamma]_1 = [\gamma]_2 = \mu^*$ and

$$R\left(\gamma(dx \times dy) \| \mu(dx)p(x, dy)\right) \leq I(\mu^*) + \varepsilon.$$

Since $[\gamma]_1 = [\gamma]_2 = \mu^*$ there is $q(x, dy)$ such that $\gamma(dx \times dy) = \mu^*(dx)q(x, dy)$ and μ^* is invariant under q. Moreover by the chain rule

$$\infty > R\left(\gamma(dx \times dy) \| \mu^*(dx)p(x, dy)\right) = \int_S R\left(q(x, \cdot) \| p(x, \cdot)\right) \mu^*(dx).$$

If q were ergodic, we could use it to define controls for the representation

via $\bar{\mu}_i^n(\cdot) = q(\bar{X}_{i-1}^n, \cdot)$, and then by the L^1-ergodic theorem

$$\begin{aligned}
&\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} \\
&= \limsup_{n\to\infty} \inf_{\{\bar{\mu}_i^n\}} E\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^n R\left(\bar{\mu}_i^n(\cdot)\,\middle\|\, p(\bar{X}_{i-1}^n, \cdot)\right)\right] \\
&\leq \limsup_{n\to\infty} E\left[F(\bar{L}^n) + \frac{1}{n}\sum_{i=1}^n R\left(q(\bar{X}_{i-1}^n, \cdot)\,\middle\|\, p(\bar{X}_{i-1}^n, \cdot)\right)\right] \\
&= \lim_{n\to\infty} E\left[F(\bar{L}^n) + \int_S R\left(q(x,\cdot)\,\|p(x,\cdot)\right)\bar{L}^n(dx)\right] \\
&= \left[F(\mu^*) + \int_S R\left(q(x,\cdot)\,\|p(x,\cdot)\right)\mu^*(dx)\right] \\
&\leq \inf_{\mu}\left[F(\mu) + I(\mu)\right] + 2\varepsilon,
\end{aligned}$$

and since $\varepsilon > 0$ is arbitrary we would be done.

The problem is that we don't know $q(x, dy)$ is ergodic. To deal with this we add a little bit of p to q to make the combination ergodic. For $\delta \in (0,1)$ let $\mu^\delta = (1-\delta)\mu^* + \delta\pi$. Since F is continuous we can choose $\delta > 0$ so that $F(\mu^\delta) \leq F(\mu^*) + \varepsilon$, and by convexity one also has

$$I(\mu^\delta) = I((1-\delta)\mu^* + \delta\pi) \leq (1-\delta)I(\mu^*) + \delta I(\pi) = (1-\delta)I(\mu^*)$$

since $I(\pi) = 0$. Let

$$\gamma^\delta(dx \times dy) = (1-\delta)\mu^*(dx)q(x, dy) + \delta\pi(dx)p(x, dy).$$

Then it is easy to check that $[\gamma^\delta]_1 = [\gamma^\delta]_2 = (1-\delta)\mu^* + \delta\pi$, and therefore there is a transition kernel $q^\delta(x, dy)$ such that $\gamma^\delta(dx \times dy) = \mu^\delta(dx)q^\delta(x, dy)$. It is not hard to show that $q^\delta(x, dy)$ inherits the same transitivity condition (30) [14, Lemma 8.6.3], and therefore q^δ is ergodic with unique stationary distribution μ^δ. Thus

$$\limsup_{n\to\infty} -\frac{1}{n}\log Ee^{-nF(L^n)} \leq \inf_{\mu}\left[F(\mu) + I(\mu)\right] + 2\varepsilon,$$

and the result follows.

Finally we remark on the uniformity. The large deviation result just stated applies to each initial condition separately, and in principle "how large n has to be" could depend on x_0. However, when combined with the Markov property the last condition of the theorem gives uniformity in an open neighborhood of each x_0. Since S is compact, an open covering argument gives global uniformity.

Lecture 12: Current Developments and Related Problems

The use of large deviation ideas in the design and analysis of Monte Carlo schemes is just beginning. In this final lecture we make some remarks on areas where the insights of an appropriate large deviations analysis may make a difference with regard to whether or not one can effectively evaluate probabilities and expected values involving rare events.

In order to keep the presentation focused we have emphasized the theory and computational methods associated with "light-tailed" random variables. There is also significant interest when the "driving noises" of the system have heavy tails, though there are far fewer general results to date and the models and dynamics are typically much more constrained than in the light tailed setting. A good reference for some theoretical results is [37], and references on importance sampling include [38, 2, 20].

1 Homogenization and problems with multiple scales

Many problems in the physical sciences involve multiple temporal and/or spatial scales. An example of a potential energy surface with multiple scales (a "rough" energy landscape) is illustrated in Figure 19. The left hand

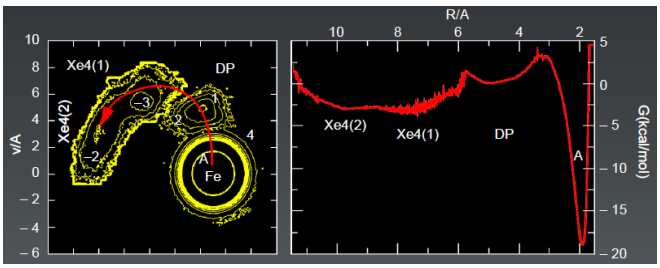

Figure 19: A rough energy landscape

panel plots level curves of the energy in two dimensions. The right hand side follows the energy U as a particle traces out the red arrow moving from the end of the arrow to its base. The relevant process model is

$$dX^\varepsilon = -\nabla U(X^\varepsilon)dt + \sqrt{\varepsilon}dW,$$

and one is interested in the probability of a transition from a neighborhood of the end of the arrow to the deep basin of attraction at the beginning of the arrow.

To model such a situation one might consider $U(x)$ to be the sum of a smoothly varying component $V(x)$ plus $\varepsilon Q(x/\delta, \gamma)/\delta$, where $Q(\cdot, \gamma)$ is an

ergodic random field, and both $\varepsilon > 0$ and $\delta > 0$ are "small," and generalizations such as $\varepsilon F(x)Q(x/\delta, \gamma)/\delta$ with F deterministic. Based on considerations from the application, one should assume $\varepsilon/\delta \to \infty$ as they tend to zero. A natural choice would be to use Gaussian random fields.

The large deviation theory in the form one would need for to applications to Monte Carlo has not been established for these problems (though one can identify algorithms that seem to work well be analogy with simpler problems, such as when $Q(\cdot)$ is periodic but not random). One issue is that the LLN problem is still not well understood for either the controlled or uncontrolled version of the diffusion model. References on the LDP for periodic coefficients include [1, 26, 33]. Two papers which consider some aspects of large deviations for random coefficients are [40, 41]. However, these papers do not give an explicit form for the rate function, and hence are not suitable as a basis for the design of Monte Carlo. A paper which suggests an importance sampling scheme for the case of random media by analogy with the case of periodic media is [27].

2 Simulation near rest points

The problems considered in Lectures 9 and 10 for applications of importance sampling involved either a risk-sensitive cost or a probability of escape over a finite time interval, or escape from a set that did not include a rest point of the noiseless dynamics. In Lecture 5 we briefly commented on the Freidlin-Wentsell theory, which allows one to tie together large deviation estimates over finite time intervals to prove results for the escape time and escape location from a domain which does include a rest point. Just as the situation when the domain contains a rest point complicated the large deviation analysis, it also complicates the construction of effective Monte Carlo methods. In this section we comment briefly on some of the difficulties.

Consider the assumptions on the process $\{X^n, n \in \mathbb{N}\}$ and domain G that were made in Lecture 5. Thus we assumed that on any finite time interval $[0, T]$, $\{X^n, n \in \mathbb{N}\}$ satisfies an LDP with rate I_T that is uniform in the initial condition in compact sets, that the rate function vanishes on the solutions to $\dot{\phi} = b(\phi)$, and that $\langle b(x), n(x) \rangle < 0$ for all $x \in \partial G$. The origin was assumed asymptotically stable under $\dot{\phi} = b(\phi)$, and we also assumed a nondegeneracy condition on the function L appearing in the definition of I_T that was sufficient but not necessary for the results stated in Lecture 5.

Among the quantities one would like to compute more accurately are the mean escape time $E_0\tau^n$, where $\tau^n \doteq \{t : X^n(t) \notin G\}$. Under the conditions just stated we know

$$n \log E_0\tau^n \to \inf \{Q(z) : z \in \partial G\},$$

where Q is the quasipotential. A more accurate approximation could, in principle, be obtained through Monte Carlo. However, unlike the problems

of estimating an exponentially small probability discussed in Lectures 9 and 10, no straightforward way to speed up the estimation of $E_0\tau^n$ has emerged. The difficulties are not so much due to the relative variance of the standard Monte Carlo estimator, but rather due to the fact that the construction of even a single sample will scale exponentially in n. When estimating a small probability the random variable of interest is supported on two points (i.e., 0 and 1) which are far apart relative to the quantity being estimated, and for this reason the variance is large. The distribution of τ^n is peaked near $\exp n \inf\{Q(z) : z \in \partial G\}$, and for this reason samples will scale exponentially in n, and any standard form of importance sampling (or other scheme such as standard splitting) does not change this in any meaningful way.

Among the surrogates that one might consider for $E_0\tau^n$ is $P_0\{\tau^n \leq T\}$, where T is large but order 1 (i.e., it does not scale with n). If one can construct a subsolution to the corresponding PDE

$$V_t(x,t) + \mathbb{H}(x, DV(x,t)) = 0,$$

$$V(x,t) = 0 \text{ for } x \in \partial G \times [0,T) \text{ and } V(x,T) = \infty \text{ for } x \in G$$

then the theory of Lectures 9 and 10 applies and one obtains good performance. However, in many problems one is tempted to exploit the fact that T is large, and therefore seek a *time-independent* subsolution, i.e., a function $\bar{V}(x)$ that satisfies

$$\mathbb{H}(x, D\bar{V}(x)) \geq 0 \text{ and } \bar{V}(x) \leq 0 \text{ for } x \in \partial G, \tag{32}$$

and with value $\bar{V}(0) \leq V(0,0)$ that is close to $V(0,0)$. For example, with an appropriate choice of the constant c, one can show that $\bar{V}(x) = -Q(x) + c$ satisfies (32), and that $V(0,0) \to \bar{V}(0)$ as $T \to \infty$.

However it turns out, as remarked previously, that the use of subsolutions is more subtle when the domain contains a rest point. In particular, straightforward use of the scheme generated by $\bar{V}$ that satisfied (32) will lead to a scheme that degrades sharply as $T \to \infty$. The problem is related to the fact that when a rest point is present one can construct subsolutions, but not *strict* subsolutions near the rest point (a strict subsolution would satisfy $\mathbb{H}(x, D\bar{V}(x)) \geq \delta$ for some $\delta > 0$). Because of this, simulated trajectories that remain in a neighborhood of the origin build up likelihood ratios that are exponentially large in T, and thus lead to large variance. These trajectories are in fact similar to the "rogue" trajectories mentioned in Lecture 8. A detailed discussion of the issues, as well as one possible approach to dealing with the difficulties (but only in dimension 1!) can be found in [28].

3 Infinite dimensional systems and SPDEs

Infinite dimensional stochastic systems in continuous time take many forms. Examples familiar to the author of these notes include stochastic flows of diffeomorphisms [17, 8], stochastic partial differential equations driven by Poisson noise [4], and weakly interacting systems of many particles [6]. In the first example a system driven by infinite dimensional Brownian motion is used to construct a random diffeomorphism (change of variable), which is then to define a prior for a Bayesian approach to the problem of matching shapes (specifically three dimensional shapes that arise in medical imaging). The change of variable can be thought of as a random "rubber sheet", and any model driven by a finite collection of Brownian motions would artificially favor certain distortions of the sheet. The second example arises in modeling the flow of pollutants in an aquifer or other body of water, where the pollutant enters via discrete injections (leading to the Poisson driving noise), and drives a linear PDE used to account for advection and diffusion. The final example is essentially infinite dimensional, since one is interested in a large number of particles, and studies the system as the number of particles tends to infinity.

The approach based on representations and weak convergence is especially attractive for proving large deviation estimates for infinite dimensional systems, since it does not use discretizations or approximations of any sort, nor does it require uniqueness for an infinite dimensional nonlinear partial differential equation. Indeed, a large number of authors have used the representation for infinite dimensional Brownian motion to study many different types of models, including stochastic wave equations, stochastic Navier-Stokes, stochastic Volterra equation, etc. As mentioned in Lecture 4, a representation for functionals of a general Poisson random measure on a Polish space, and also for functionals of an infinite dimensional Brownian motion and an independent Poisson random measure were recently established. Thus one has representations for many of the processes one might encounter driven by infinitely divisible noises.

With useful representations available and certain a priori constructions that allow one to restrict to controls that take values in a compact set, the general approach usually follows the arguments used in the simple case of stochastic differential equations treated in Lecture 4. Thus one needs a strong solution, by which we mean a solution that leads to a measurable mapping from the noise space to the system output, as well as a good understanding of qualitative properties such as tightness for the controlled version of the original model and LLN limits. As mentioned previously there is great variety to the problem and model formulations in the infinite dimensional setting, and these basic qualitative properties that one needs to carry out the analysis require a fairly deep understanding of the nuances of the specific problem formulation.

As a last comment, it should be noted that there is very little experience with the design and use of importance sampling in the infinite dimensional setting. In this setting the construction of even a single sample is usually expensive, and so there is even more value to efficient simulation methods. The cost of constructing samples suggests that speedup would be useful even for events that are not so far out in the tail of the distribution, and so the moderate deviation theory for these problems may be useful.

References

[1] P. Baldi. Large deviations for diffusions processes with homogenization and applications. *Annals of Probability*, 10:509–524, 1991.

[2] J.H. Blanchet, P. Glynn, and J.C. Liu. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *QUESTA*, 57:99–113, 2007.

[3] M. Boué and P. Dupuis. A variational representation for certain functionals of Brownian motion. *The Annals of Prob.*, 26:1641–1659, 1998.

[4] A. Budhiraja, J. Chen, and P. Dupuis. Large deviations for stochastic partial differential equations driven by a Poisson random measure. *Stoch. Proc. and Appl.*, 123:523–560, 2013.

[5] A. Budhiraja and P. Dupuis. A variational representation for positive functionals of infinite dimensional brownian motion. *Prob. Math. Statist.*, 20:39–61, 2000.

[6] A. Budhiraja, P. Dupuis, and M. Fischer. Large deviation properties of weakly interacting processes via weak convergence methods. *Ann. Probab.*, 40:74–1–2, 2012.

[7] A. Budhiraja, P. Dupuis, and V. Maroulas. Large deviations for infinite dimensional stochastic dynamical systems. *Ann. Probab.*, 36:1390–1420, 2008.

[8] A. Budhiraja, P. Dupuis, and V. Maroulas. Large deviations for stochastic flows of diffeomorphisms. *Bernoulli Journal*, 16:234–257, 2010.

[9] A. Budhiraja, P. Dupuis, and V. Maroulas. Variational representations for continuous time processes. *Ann. de l'Inst. H. Poincaré*, 47:725–747, 2011.

[10] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviations approach to design and analysis. *Stoch. Proc. Appl.*, 119:562–587, 2009.

[11] T. Dean and P. Dupuis. The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. *Annals of OR*, 189:63–102, 2011.

[12] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.*, 28:1–47, 1975.

[13] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, III. *Comm. Pure Appl. Math.*, 29:389–461, 1976.

[14] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

[15] P. Dupuis and R.S. Ellis. Large deviations for Markov processes with discontinuous statistics, II: Random walks. *Probab. Th. Rel. Fields*, 91:153–194, 1992.

[16] P. Dupuis, R.S. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Annals of Probability*, 19:1280–1297, 1991.

[17] P. Dupuis, U. Grenander, and M. Miller. A variational formulation of a problem in image matching. *Quarterly of Applied Mathematics*, 56:587–600, 1998.

[18] P. Dupuis and H. J. Kushner. Large deviations estimates for systems with small noise effects, and applications to stochastic systems theory. *SIAM J. Control Optimization*, 24:979–1008, 1986.

[19] P. Dupuis and H. J. Kushner. Stochastic systems with small noise, analysis and simulation; a phase locked loop example. *SIAM J. Appl. Math.*, 47:643–661, 1987.

[20] P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM Trans. Modeling Comp. Simulation*, 17:1–21, 2007.

[21] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with slow-down. *QUESTA*, 57:71–83, 2007.

[22] P. Dupuis, K. Leder, and H. Wang. On the large deviations properties of the weighted-serve-the-longest-queue policy. In V. Sidoravicius and M.E. Vares, editors, *In and Out of Equilibrium 2*. Birkhauser, New York, 2008.

[23] P. Dupuis, K. Leder, and H. Wang. Importance sampling for weighted serve-the-longest-queue. *Math. of Operations Research*, 34:642–660, 2009.

[24] P. Dupuis, Y. Liu, N. Plattner, and J.D. Doll. On the infinite swapping limit for parallel tempering. *SIAM J. Multiscale Model. Simul.*, 10:986–1022, 2012.

[25] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Ann. Appl. Prob.*, pages 1306–1346, 2007.

[26] P. Dupuis and K. Spiliopoulos. Large deviations for multiscale diffusions via weak convergence methods. *Stoch. Proc. and Their Appl.*, 122:1947–1987, 2012.

[27] P. Dupuis, K. Spiliopoulos, and H. Wang. Importance sampling for multiscale diffusions. *SIAM J. Multiscale Model. and Simul.*, 10:1–27, 2012.

[28] P. Dupuis, K. Spiliopoulos, and X. Zhou. Escaping from an attractor: Importance sampling and rest points. *preprint*, 2012.

[29] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.

[30] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62:113–157, 2009.

[31] P. Dupuis and J. Zhang. Explicit solutions for a class of nonlinear PDE that arise in allocation problems. *SIAM J. on Mathematical Analysis*, pages 1627–1667, 2008.

[32] A. Figotin, A. Gordon, S. Molchanov, J.Quinn, and N. Stavrakas. Occupancy numbers in testing random number generators. *SIAM J. Appl. Math*, 62:1980–2011, 2002.

[33] M. I. Freidlin and R. Sowers. A comparison of homogenization and large deviations, with applications to wavefront propagation. *Stoch. Proc. and Their Appl.*, 82:23–52, 1991.

[34] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, New York, 1984.

[35] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer–Verlag, New York, 2004.

[36] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746, 1997.

[37] H. Hult, F. Lindskog, T. Mikosch, and G. Samorodnitsky. Functional large deviations for multivariate regularly varying random walks. *Ann. Appl. Prob.*, 15:2651–2680, 2005.

[38] H. Hult and J. Svensson. On importance sampling with mixtures for random walks with heavy tails. *TOMACS*, 22:1–21, 2012.

[39] R.V. Kohn, F. Otto, M.G. Reznikoff, and E. Vanden-Eijnden. Action minimization and sharp interface limits for the stochastic Allen-Cahn equation. *Comm. Pure Appl. Math.*, 60:393–438, 2007.

[40] E. Kosygina, F. Rezakhanlou, and S. R. S. Varadhan. Stochastic homogenization of Hamilton-Jacobi-Bellman equations. *Comm. Pure Appl. Math.*, 59:1–33, 2006.

[41] P.-L. Lions and P.E. Souganidis. Homogenization of viscous Hamilton-Jacobi-Bellman equations in stationary ergodic media. *Communications in Partial Differential Equations*, 30:335–375, 2006.

[42] R.O. Moore, G. Biondini, and W.L. Kath. Importance sampling for noise-induced amplitude and timing jitter in soliton transmission systems. *Optics Letters*, 28:105–107, 2003.

[43] P. Dai Pra, L. Meneghini, and W. J. Runggaldier. Some connections between stochastic control and dynamic games. *MCSS*, 9:303–326, 1996.

[44] Z. Schuss. *Nonlinear Filtering and Optimal Phase Tracking*. Springer, New York, 2011.

[45] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing*. Chapman and Hall, New York, 1995.

[46] E. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer-Verlag, New York, 1998.

[47] P. Whiting, C. Nuzman, and J. Zhang. Importance sampling and the design of an optical switch. In *Allerton Conference on Communications, Computing and Control*, Urbana, Illinois, 2002.

[48] J. Zhang and P. Dupuis. Large deviation principle for general occupancy models. *Combinatorics, Probability, and Computing*, 17:437–470, 2008.