Stochastic Process

A Brief Note about Information, Markovian Process, and Least-Action Principle

Table of contents

Prologue

This is a little book about stochastic process. We start with the axiomatic system of information (chapter 1). Then using information, we introduce the continuous time Markovian process, and show how it relax to equilibrium (chapter 2). We then move on to add spatial smoothness to Markovian process, which results in many interesting results including Langevin process (chapter 3). In the end, we generalize least-action principle to distribution and comparing data-fitting with least-action principle (chapter 4). Readers may omit the sections in which the titles start with ♣. They are interesting digressions. Also the sections start with ♡, which contains the materials for future investigation.

The mathematical techniques employed here will not go beyond the basic calculus (Taylor expansion, improper integral, and integration by parts) and linear algebra (equivalence relation and equivalence class, matrix manipulations, orthogonal diagonalization, and determinant). Knowing the basic probability theory (normal distribution and Gaussian integral) will be beneficial. We try to make it self-contained, and introduce new concept or technique only when it is essential. Statements like "obviously…" and "apparently…" are avoided; and we try to display all the steps of calculation without omitting any of them.

For each section, the title is a sentence that briefly summarizes the whole section. We use bold font for **definition** and italic font for *emphasis*. Only important equations are numbered. So, readers can quickly skim through by reviewing section titles, bold and italic texts, and numbered equations.

At last, this book is written by TEXmacs, using the GFDL-1.3 license. You can contact me with e-mail: shuiruge@whu.edu.cn.

Chapter 1
Relative Entropy

In this chapter, we discuss the axiomatization of information. By figuring out how Shannon entropy fails for continuous random variable, we build a proper axiomatic system on which the expression of information is found to be unique.

1.1 Conventions in This Chapter

Those that are not deterministic are denoted by capital letters. But, a capital letter may also denote something that is determined. For example, a random variable has to be denoted by capital letter, like X, while we can also use F to denote something determined, such as a functional.

1.2 A Brief Review of Probability

Maybe the best analogy of probability is deterministic. Consider a pile of sand distributed on a table. At each location of the table, there is a density of sands. And for each area of the table, we can compute the total mass of sands in that area. When we move some portion of sands from one area to another on the table, the mass of all sands keeps constant. Even though the pile of sand model lacks any randomness, it does furnish a visual picture for the main concepts that constitute probability.

The set of all possible values of a random variable is called the **alphabet** (a synonym for table).[1.1] And for each value in the alphabet, we assign a *positive* value called **density** if the alphabet is of continuum (continuous random variable), or **mass** otherwise (discrete random variable).[1.2] We use **distribution** for not only the mass or density on the alphabet, but also a sampler that can sample an ensemble of values of the random variable that converges to the mass or density when the number of sample tends to infinity. For example, we say X is a random variable with alphabet $\mathcal{X}$ and distribution P.

The density of a value x is usually denoted by $p(x)$, which, as a function, is called **density function**. Notice that $p(x)$ is deterministic, thus not capital. The same for mass, where $p(x)$ is called **mass function**. Thus, we can say the expectation of a function f on distribution P, denoted by $\mathbb{E}_P[f]$ or $\mathbb{E}_{x\sim P}[f(x)]$. If the alphabet $\mathcal{X}$ is of continuum, then it is $\int_{\mathcal{X}} \mathrm{d}x p(x)\, f(x)$, otherwise $\sum_{x\in\mathcal{X}} p(x)\, f(x)$.

If there exists random variables Y and Z, with alphabets $\mathcal{Y}$ and $\mathcal{Z}$ respectively, such that $X = Y \oplus Z$ (for example, let X two-dimensional, Y and Z are the components), then we have **marginal distribution**s, denoted by P_Y and P_Z, where $p(y) := \int_{\mathcal{Z}} \mathrm{d}z p(y,z)$ and $p(z) := \int_{\mathcal{Y}} \mathrm{d}y p(y, z)$ if X is of continuum, and the same for mass function. Notice that we have omitted the subscript Y in p_Y (and the same for p_Z) since the y in $p(y)$ has clearly indicated this. We **marginalize** Z so as to get P_Y.

1.1. Many textures call it **sample space**. But "space" usually hints for extra structures such as vector space or topological space. So, we use "alphabet" instead, following David Mackay. (See his remarkable book *Information Theory, Inference, and Learning Algorithms*, section 2.1. Link to free PDF: https://www.inference.org.uk/itprnn/book.pdf)

1.2. In many textures, the density or mass function is non-negative (rather than being positive). Being positive is beneficial because, for example, we will discuss the logarithm of density or mass function, for which being zero is invalid. For any value on which density or mass function vanishes, we throw it out of $\mathcal{X}$, which in turn guarantees the positivity.

We further have the **conditional distribution** of Y given Z, denoted by $P_{Y|Z}$, where $p(y|z) := p(y,z)/p(z)$ (we omit the subscript of $p_{Y|Z}$ too). Suppose that we samples lots of (Y,Z) values from P, and then filters the pairs with $Z=z$. The frequency of $Y=y$ found in the filtered samples is approximated by $p(y|z)$.

1.3 Shannon Entropy Is Plausible for Discrete Random Variable

The Shannon entropy is well-defined for discrete random variable. Let X a discrete random variables with alphabet $\{1,\dots,n\}$ with p_i the mass of $X=i$. The Shannon entropy is thus a function of $(p_1,\dots,p_n)$ defined by

$$H(P) := -k\sum_{i=1}^{n} p_i \ln p_i,$$

where k is a positive constant. Interestingly, this expression is unique given some plausible axioms, which can be qualitatively expressed as

1. H is a continuous function of $(p_1,\dots,p_n)$;
2. larger alphabet has higher uncertainty (information or entropy); and
3. if we have known some information, and based on this knowledge we know further, the total information shall be the sum of all that we know.

Here, we use **uncertainty**, **surprise**, **information**, and **entropy** as interchangeable.

The third axiom is also called the additivity of information. For two independent variables X and Y with distributions P and Q respectively, the third axiom indicates that the total information of $H(PQ)$ is $H(P)+H(Q)$. But, the third axiom indicates more than this. It also defines a "conditional entropy" for dealing with the situation where X and Y are dependent. Jaynes gives a detailed declaration to these axioms.[1.3] This conditional entropy is, argued by others, quite strong and not sufficiently natural. The problem is that this stronger axiom is essential for Shannon entropy to arise. Otherwise, there will be other entropy definitions that satisfy all the axioms, where the third involves only independent random variables, such as Rényi entropy.[1.4]

As we will see, when extending the alphabet to continuum, this problem naturally ceases.

1.4 Shannon Entropy Fails for Continuous Random Variable

The Shannon entropy, however, cannot be directly generalized to continuous random variable. Usually, the entropy for continuous random variable X with alphabet $\mathcal{X}$ and distribution P is given as a functional of the density function $p(x)$,

$$H(P) := -k\int_{\mathcal{X}} \mathrm{d}x\, p(x)\ln p(x)$$

which, however, is not well-defined. The first issue is that the p has dimension, indicated by $\int_{\mathcal{X}}\mathrm{d}x\, p(x)=1$. This means we put a dimensional quantity into logarithm which is invalid. The second issue is that the H is not invariant under coordinate transformation $X\to Y:=\varphi(X)$ where φ is a diffeomorphism. But as a "physical" quantity, H should be invariant under "non-physical" transformations.

To eliminate the two issues, we shall extends the axiomatic description of entropy. The key to this extension is introducing another distribution, Q, which has the same alphabet as P; and instead considering *the uncertainty (surprise) caused by P when prior knowledge has been given by Q*. As we will see, this will solve the two issues altogether. Explicitly, we extend and quantify the axioms in section 1.3 as follow.

1.3. See the appendix A of *Information Theory and Statistical Mechanics* by E. T. Jaynes, 1957. A free PDF version can be found on Internet: https://bayes.wustl.edu/etj/articles/theory.1.pdf.

1.4. *On measures of information and entropy* by Alfréd Rényi, 1961. A free PDF version can be found on Internet: http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf.

Axiom 1.1. *Given distributions P and Q on the same alphabet, H is the uncertainty caused by P when Q is known, satisfying the following conditions:*

1. *H is a smooth and local functional of p and q;*
2. *$H(P,Q)>0$ with $P\neq Q$ and $H(P,P)=0$; and*
3. *If $X=Y\oplus Z$, and if Y and Z independent, then $H(P,Q)=H(P_Y,Q_Y)+H(P_Z,Q_Z)$, where $P_Y,\dots,Q_Z$ are marginal distributions.*

The first axiom employs the locality of H, which is thought as natural since H has been a functional. The second axiom indicates that H vanishes only when there is no surprise caused by P (thus $P=Q$). It is a little like the second axiom for Shannon entropy. The third axiom, like the third in Shannon entropy, claims the additivity of surprise: if X has two independent parts, the total surprise shall be the sum of each.

1.5 Relative Entropy Is the Unique Solution to the Axiom

We are to derive the explicit expression of H based on the three axioms. The result is found to be unique.

Based on the first axiom, there is a function $h\colon(0,+\infty)\times(0,+\infty)\to[0,+\infty)$ such that H can be expressed as

$$H(P,Q)=\int_{\mathcal{X}}\mathrm{d}x\,p(x)\,h(p(x),q(x)).$$

We are to determine the explicit form of h. Thus, from second axiom,

$$H(P,P)=\int_{\mathcal{X}}\mathrm{d}x\,p(x)\,h(p(x),p(x))=0$$

holds for all distribution P. Since p is positive and h is non-negative, then we have $h(p(x),p(x))=0$ for all $x\in\mathcal{X}$. The distribution P is arbitrary, thus we find $h(x,x)=0$ for any $x\in(0,+\infty)$.

Now come to the third axiom. Since Y and Z are independent, $H(P,Q)$ can be written as $\int_{\mathcal{X}}\mathrm{d}y\mathrm{d}z\,p_Y(y)\,p_Z(z)\,h(p_Y(y)\,p_Z(z),q_Y(y)\,q_Z(z))$. Thus, the third axiom implies

$$\int_{\mathcal{X}}\mathrm{d}y\mathrm{d}z\,p_Y(y)\,p_Z(z)[h(p_Y(y)\,p_Z(z),q_Y(y)\,q_Z(z))-h(p_Y(y),q_Y(y))-h(p_Z(z),q_Z(z))]=0.$$

Following the previous argument, we find $h(ax,by)=h(a,b)+h(x,y)$ for any $a,b,x,y\in(0,+\infty)$. Taking derivative on a and b results in $\partial_1h(ax,by)\,x=\partial_1h(a,b)$ and $\partial_2h(ax,by)\,y=\partial_2h(a,b)$. Since $\partial_1h(a,a)+\partial_2h(a,a)=(\mathrm{d}/\mathrm{d}a)\,h(a,a)=0$, we get $\partial_1h(ax,ay)\,x+\partial_2h(ax,ay)\,y=0$. Letting $a=1$, it becomes a first order partial differential equation $\partial_1h(x,y)\,x+\partial_2h(x,y)\,y=0$, which has a unique solution that $h(x\mathrm{e}^t,y\mathrm{e}^t)$ is constant for all t. Choosing $t=-\ln y$, we find $h(x,y)=h(x/y,1)$. Now h reduces from two variables to one. So, plugging this result back to $h(ax,by)=h(a,b)+h(x,y)$, we have $h(xy,1)=h(x,1)+h(y,1)$. It looks like a logarithm. We are to show that it is indeed so. By taking derivative on x and then letting $y=1$, we get an first order ordinary differential equation $\partial_1h(x,1)=\partial_1h(1,1)/x$, which has a unique solution that $h(x,1)=\partial_1h(1,1)\ln(x)+C$, where C is a constant. Combined with $h(x,y)=h(x/y,1)$, we finally arrive at $h(x,y)=\partial_1h(1,1)\ln(x/y)+C$. To determine the $\partial_1h(1,1)$ and C, we use the second axiom $\partial_1h(1,1)\int\mathrm{d}x\,p(x)\ln(p(x)/q(x))+C>0$ when $p\neq q$ and $\partial_1h(1,1)\int\mathrm{d}x\,p(x)\ln(p(x)/p(x))+C=0$. The second equation results in $C=0$. By Gibbs' inequality, the integral $\int\mathrm{d}x\,p(x)\ln(p(x)/q(x))$ is non-negative, thus from the first equation, $\partial_1h(1,1)>0$. Up to now, all things about h have been settled. We conclude that there is a unique expression that satisfies all the three axioms, which is

$$H(P,Q)=k\int_{\mathcal{X}}\mathrm{d}x\,p(x)\ln\frac{p(x)}{q(x)},$$

where $k>0$. This was first derived by Solomon Kullback and Richard Leibler in 1951, so it is called **Kullback–Leibler divergence** (**KL-divergence** for short), denoted by $D_{\mathrm{KL}}(P\|Q)$. Since it characterizes the relative surprise, it is also called **relative entropy** (entropy for surprise).

The locality is essential for relative entropy to arise. For example, Renyi divergence, defined by

$$H_\alpha(P,Q)=\frac{1}{\alpha-1}\ln\left(\int_{\mathcal{X}}\mathrm{d}x\,\frac{p^\alpha(x)}{q^{\alpha-1}(x)}\right),$$

also satisfies the three axioms when locality is absent.

In the end, we examine the two issues appeared in Shannon entropy (section 1.4). In $H(P,Q)$, the logarithm is $\ln(p/q)$ which is dimensionless. And a coordinate transformation $X\to Y:=\varphi(X)$ makes $\int\mathrm{d}x\,p(x)=\int\mathrm{d}y\,|\det(\partial\varphi^{-1})(y)|\,p(\varphi^{-1}(y))=:\int\mathrm{d}y\,\tilde{p}(y)$, thus $p\to\tilde{p}:=|\det(\partial\varphi^{-1})|\,p\circ\varphi^{-1}$. The same for $q\to\tilde{q}:=|\det(\partial\varphi^{-1})|\,q\circ\varphi^{-1}$. The common factor $|\det(\partial\varphi^{-1})|$ will be eliminated in $\ln(p/q)$, leaving $H(P,Q)$ invariant (since $\int\mathrm{d}x\,p\ln(p/q)\to\int\mathrm{d}y\,\tilde{p}\ln(\tilde{p}/\tilde{q})$, which equals to $\int\mathrm{d}x\,p\ln(p/q)$). So, the two issues of Shannon entropy cease in relative entropy.

Chapter 2

Master Equation and Detailed Balance

In this chapter, we discuss continuous time Markovian process. Using the result obtained in chapter 1, we declare how a Markovian process relaxes to its stationary equilibrium.

2.1 Conventions in This Chapter

Follow the conventions in chapter 1. Let X a multi-dimensional random variables, being, discrete, continuous, or partially discrete and partially continuous, with alphabet $\mathcal{X}$ and distribution P. Even though the discussion in this chapter applies to both discrete and continuous random variables, we use the notation of the continuous. The reason is that converting from discrete to continuous may cause problems (section 1.4), while the inverse is safe and direct.

When using conditional density function to describe the probability of transition, like $p_{t\to t'}(x'|x)$ which denotes the transition from x at time t to x' at time t', we find it convenient to adopt the notation to $p_{t\to t'}(x\to x')$.

2.2 Master Equation Describes the Evolution of Markov Process

We generalize the pile of sand model in section 1.2, imagining that these sands are magic, having free will to move on the table. The distribution of sands changes with time. In the language of probability, the density of sands at position x on the table is described by a time-dependent density function $p(x,t)$ where t represents time. The total mass of the sands is kept invariant and normalized to one. The alphabet $\mathcal{X}$ is the table itself (namely, every position on the table).

Let $q_{t\to t'}(x\to x')$ denote the *portion* of density at position x at time t that transits to position x' at time t'. Then, the transited density will be $p(x,t)q_{t\to t'}(x\to x')$. There may be some portion of density at position x that does not transit during $t\to t'$ (the lazy sands). In this case we regard the sands as transiting from position x to x (staying on x), which is $q_{t\to t'}(x\to x)$. Now, every sand at position x has transited during $t\to t'$, and the total portion shall be 100%, which means

$$\int_{\mathcal{X}} \mathrm{d}x'\, q_{t\to t'}(x\to x') = 1. \tag{2.1}$$

As portion, $q_{t\to t'}$ cannot be negative, thus $q_{t\to t'}(x\to x')\geqslant 0$ for each x and x' in $\mathcal{X}$. We call $q_{t\to t'}$ the **transition density**. Not like the density function of distribution, transition density can be zero in a subset of $\mathcal{X}$.

The transition makes a difference in the density at position x'. The difference is caused by the density transited from x', which is $\int_{\mathcal{X}} \mathrm{d}x p(x',t) q_{t\to t'}(x'\to x)$, and that transited to x', which is $\int_{\mathcal{X}} \mathrm{d}x p(x,t) q_{t\to t'}(x\to x')$. Thus, we have

$$p(x',t') - p(x',t) = \int_{\mathcal{X}} \mathrm{d}x \left[p(x,t) q_{t\to t'}(x\to x') - p(x',t) q_{t\to t'}(x'\to x)\right].$$

By inserting equation 2.1, we find

$$p(x',t') = \int_{\mathcal{X}} \mathrm{d}x p(x,t) q_{t\to t'}(x\to x'), \tag{2.2}$$

which is called the **discrete time master equation**. When $t'=t$, we have $p(x',t) = \int_{\mathcal{X}} \mathrm{d}x p(x,t) q_{t\to t}(x\to x')$, indicating that

$$q_{t\to t}(x\to x') = \delta(x-x'),$$

where $\delta(x-x')$ indicates Kronecker's delta $\delta_{x,x'}$ when $\mathcal{X}$ is discrete, or Dirac's delta function when $\mathcal{X}$ is continuous. Even though we call it Dirac's delta function, it is in fact a generalized function. A generalized function has meaning only when it is integrated together with other (not generalized) functions. For example, Dirac's delta function has $\int_{\mathcal{X}} \mathrm{d}x \delta(x-y) f(x) = f(y)$ for any usual function f.

In addition, if the change of the distribution of sands is smooth, that is, there is not a sand lump that jumping from one place to another in an arbitrarily short period of time, then $q_{t\to t'}$ is smooth on t'. Taking derivative on t' and then setting t' to t, we have

$$\frac{\partial p}{\partial t}(x',t) = \int_{\mathcal{X}} \mathrm{d}x p(x,t)\, r_t(x,x'), \tag{2.3}$$

where

$$r_t(x,x') := \lim_{t'\to t} \frac{\partial q_{t\to t'}}{\partial t'}(x\to x')$$

is **transition rate**. Equation 2.3 is called the **continuous time master equation**, or simply **master equation**. The word "master" indicates that the transition rate has completely determined (mastered) the evolutionary behavior of distribution.

Even though all these concepts are born of the pile of sand, they are applicable to any stochastic process where the distribution $P(t)$ is time-dependent (but the alphabet $\mathcal{X}$ is time-invariant), no matter whether the random variable is discrete or continuous.

A stochastic process is **Markovian** if the transition density $q_{t\to t'}$ depends only on the time difference $\Delta t := t'-t$, thus $q_{t\to t'}$ is re-denoted by $q_{\Delta t}$. In this situation, transition rate r becomes time-independent, so the master equation turns to be

$$\frac{\partial p}{\partial t}(x',t) = \int_{\mathcal{X}} \mathrm{d}x p(x,t)\, r(x,x'). \tag{2.4}$$

Since we only deal with Markovian stochastic process throughout this note, when referring to master equation, we mean equation 2.4. And to discrete time master equation, we mean:

$$p(x',t+\Delta t) = \int_{\mathcal{X}} \mathrm{d}x p(x,t) q_{\Delta t}(x\to x'). \tag{2.5}$$

Before finishing this section, we discuss the demanded conditions for transition rate. The normalization of transition density 2.1 implies that $\int_{\mathcal{X}} \mathrm{d}y r(x,y) = 0$. This can be seen by Taylor expanding $q_{\Delta t}$ as $q_{\Delta t}(x\to y) = \delta(x-y) + r(x,y)\,\Delta t + o(\Delta t)$, where we have inserted $q_0(x\to y) = \delta(x-y)$ and the definition of r. Also from this Taylor expansion, we see that the non-negativity of $q_{\Delta t}$ implies $r(x,y) \geqslant 0$ when $x \neq y$. Since p is a density function of distribution, and density

function is defined to be positive (see section 1.2), equation 2.2 must conserve this positivity. We are to show that this is guaranteed by master equation itself, without any extra condition demanded for transition rate. It is convenient to use discrete notations, thus we replace $x \to i$, $y \to j$, and $\int \mathrm{d}x \to \sum$. Master equation turns to be $(\mathrm{d}p_j/\mathrm{d}t)(t) = \sum_i p_i(t)\, r_{ij}$. It is an ordinary differential equation. Recall that $r_{ij} \geqslant 0$ when $i \neq j$, and thus $r_{ii} \leqslant 0$ (since $\sum_j r_{ji} = 0$). We separate the right hand side to be $p_j(t)\, r_{jj} + \sum_{i:i\neq j} p_i(t)\, r_{ij}$, and the worst situation is that $r_{ij} = 0$ for each $i \neq j$ and $r_{jj} < 0$. In this case, master equation reduces to $(\mathrm{d}p_j/\mathrm{d}t)(t) = p_j(t)\, r_{jj}$, which has the solution $p_j(t) = p_j(0)\exp(r_{jj}t)$. It implies that $p_j(t) > 0$ as long as $p_j(0) > 0$, indicating that master equation conserves the positivity of density function. As a summary, we demand transition rate r to be $r(x,y) \geqslant 0$ when $x \neq y$ and $\int_{\mathcal{X}} \mathrm{d}y\, r(x,y) = 0$.

2.3 Transition Rate Determines Transition Density

We wonder, given a transition rate, can we obtain the corresponding transition density? Generally, we cannot get the global (or the finite) from the local (or the infinitesimal). For example, we cannot determine a function only by its first derivative at the origin. But, master equation has a group-like structure, by which the local accumulates to be global. We are to show how this happens.

We can use the master equation 2.4 to calculate $\partial^n p/\partial t^n$ for any n. For $n = 2$, by inserting master equation 2.4 (to the blue term), we have

$$\frac{\partial^2 p}{\partial t^2}(x,t) = \frac{\partial}{\partial t}\frac{\partial p}{\partial t}(x,t) = \frac{\partial}{\partial t}\int_{\mathcal{X}} \mathrm{d}x_1\, p(x_1,t) r(x_1,x) = \int_{\mathcal{X}} \mathrm{d}x_1\, \frac{\partial p}{\partial t}(x_1,t) r(x_1,x).$$

We then insert master equation 2.4 again (to the green term), and find

$$\frac{\partial^2 p}{\partial t^2}(x,t) = \int_{\mathcal{X}} \mathrm{d}x_1 \int_{\mathcal{X}} \mathrm{d}x_0\, p(x_0,t) r(x_0,x_1) r(x_1,x) = \int_{\mathcal{X}} \mathrm{d}x_0 \int_{\mathcal{X}} \mathrm{d}x_1\, p(x_0,t)\, r(x_0,x_1)\, r(x_1,x).$$

Following the same steps, it can be generalized to higher order derivatives, as

$$\frac{\partial^n p}{\partial t^n}(x,t) = \int_{\mathcal{X}} \mathrm{d}x_0 \cdots \int_{\mathcal{X}} \mathrm{d}x_{n-1}\, p(x_0,t) r(x_0,x_1) \cdots r(x_{n-1},x).$$

Notice the pattern: a leftmost $p(x_0,t)$ a sequence of r. *The reason for this pattern to arise is that transition rate r, is independent of t: a Markovian property.*

On the other hand, Taylor expand the both sides of discrete time master equation 2.5 by Δt gives, at $(\Delta t)^n$ order,

$$\frac{\partial^n p}{\partial t^n}(x,t) = \int_{\mathcal{X}} \mathrm{d}x_0\, p(x_0,t)\, q_0^{(n)}(x_0 \to x),$$

where, for simplifying notation, we have denoted the nth-order derivatives of $q_{\Delta t}$ by

$$q_{\Delta t}^{(n)}(x \to y) := \lim_{\tau \to \Delta t} \frac{\partial^n q_\tau}{\partial \tau^n}(x \to y).$$

So, by equaling the two expressions of $(\partial^n p/\partial t^n)(x,t)$, we find

$$\int_{\mathcal{X}} \mathrm{d}x_0\, p(x_0,t) \left[q_0^{(n)}(x_0 \to x) - \int_{\mathcal{X}} \mathrm{d}x_1 \cdots \int_{\mathcal{X}} \mathrm{d}x_{n-1} r(x_0,x_1) \cdots r(x_{n-1},x) \right] = 0$$

for any $n \in \{2,3,\dots\}$. This holds for all p, thus

$$q_0^{(n)}(x_0 \to x) = \int_{\mathcal{X}} \mathrm{d}x_1 \cdots \int_{\mathcal{X}} \mathrm{d}x_{n-1} r(x_0,x_1) \cdots r(x_{n-1},x).$$

Recalling that $q_{\Delta t}(x\to x') = \delta(x-x') + r(x,x')\,\Delta t + o(\Delta t)$, we have the Taylor expansion of $q_{\Delta t}$, as[2.1]

$$\begin{aligned} q_{\Delta t}(x\to x') = {} & \delta(x-x') \\ & + (\Delta t)\, r(x,x') \\ & + \frac{(\Delta t)^2}{2!}\int_{\mathcal{X}} \mathrm{d}x_1\, r(x,x_1)\, r(x_1,x') \\ & + \cdots \\ & + \frac{(\Delta t)^n}{n!}\int_{\mathcal{X}}\mathrm{d}x_1 \cdots \int_{\mathcal{X}} \mathrm{d}x_{n-1}\, r(x,x_1)\cdots r(x_{n-1},x') \\ & + \cdots. \end{aligned} \tag{2.6}$$

Well, this is a complicated formula, but its implication is straight forward and very impressive: *the transition density is equivalent to transition rate, even though transition rate is derived from the transition density with infinitesimal time-interval.*

This may be a little weird at the first sight. For example, consider $q'_{\Delta t}(x\to x') := q_{\Delta t}(x\to x') + f(x,x')\,\Delta t^2$, where f is any function ensuring that $q'_{\Delta t}$ is non-negative and normalized (thus $\int_{\mathcal{X}} \mathrm{d}y f(x,y) = 0$). Following the previous derivation, we find the discrete time master equation

$$p(x', t+\Delta t) = \int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\, q'_{\Delta t}(x\to x')$$

also leads to (continuous time) master equation 2.4 with the same r as that of $q_{\Delta t}$. So, we should have $q'_{\Delta t} = q_{\Delta t}$, which means f is not free, but should vanish.

The answer to this question is that, a transition density is not free to choose, but sharing the same degree of freedom as that of its transition rate. *The fundamental quantity that describes the evolution of a continuous time Markov process is transition rate.* For example, consider $p(z, t+\Delta t+\Delta t')$ for any Δt and $\Delta t'$. Directly, we have

$$p(z, t+\Delta t+\Delta t') = \int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\, q_{\Delta t+\Delta t'}(x\to z),$$

but on the other hand, by applying discrete time master equation twice, we find

$$\begin{aligned} p(z, t+\Delta t+\Delta t') &= \int_{\mathcal{X}} \mathrm{d}y\, p(y, t+\Delta t')\, q_{\Delta t}(y\to z) \\ &= \int_{\mathcal{X}} \mathrm{d}x\, p(x,t) \int_{\mathcal{X}} \mathrm{d}y\, q_{\Delta t}(x\to y)\, q_{\Delta t'}(y\to z). \end{aligned}$$

By equaling the two expressions of $p(z, t+\Delta t+\Delta t')$, we find

$$\int_{\mathcal{X}} \mathrm{d}x\, p(x,t) \left[q_{\Delta t+\Delta t'}(x\to z) - \int_{\mathcal{X}} \mathrm{d}y\, q_{\Delta t}(x\to y)\, q_{\Delta t'}(y\to z) \right] = 0.$$

Since p can be arbitrary, we arrive at

$$q_{\Delta t+\Delta t'}(x\to z) = \int_{\mathcal{X}} \mathrm{d}y\, q_{\Delta t}(x\to y)\, q_{\Delta t'}(y\to z). \tag{2.7}$$

This provides an addition restriction to the transition density.

2.1. Another derivation uses exponential mapping. By regarding p a time-dependent element in functional space, and r as a linear operator, it becomes (we add a hat for indicating operator, using dot $\cdot$ for its operation)

$$\frac{\mathrm{d}p}{\mathrm{d}t}(t) = \hat{r}\cdot p(t).$$

This operator differential equation has a famous solution, called exponential mapping, $p(t) = \exp(\hat{r}\, t)\, p(0)$, where the exponential operator is defined by Taylor expansion $\exp(\hat{L}) := \hat{1} + \hat{L} + (1/2!)\,\hat{L}^2 + \cdots$ for any linear operator $\hat{L}$. Indeed, by taking derivative on t on both sides, we find $(\mathrm{d}p/\mathrm{d}t)(t) = \hat{r}\cdot\exp(\hat{r}\, t)\, p(0) = \hat{r}\cdot p(t)$. Recall the discrete time master equation, $p(\Delta t) = \hat{q}_{\Delta t}\cdot p(0)$, where the transition density $\hat{q}_{\Delta t}$ is regarded as a linear operator too (so we put a hat on it). We find $\exp(\hat{r}\,\Delta t)\cdot p(0) = \hat{q}_{\Delta t}\cdot p(0)$, which holds for arbitrary $p(0)$, implying $\hat{q}_{\Delta t} = \exp(\hat{r}\,\Delta t) = 1 + \hat{r}\,\Delta t + (1/2!)\,(\hat{r}\cdot\hat{r})\,(\Delta t)^2 + \cdots$. Going back to functional representation, we have the correspondences $\hat{q}_{\Delta t}\to q_{\Delta t}(z|x)$, $\hat{r}\to r(z,x)$, $\hat{r}\cdot\hat{r}\to\int \mathrm{d}y\, r(z,y)\, r(y,x)$, $\hat{r}\cdot\hat{r}\cdot\hat{r}\to\int \mathrm{d}y_1\,\mathrm{d}y_2\, r(z,y_2)\, r(y_2,y_1)\, r(y_1,x)$, and so on, thus recover the relation between $q_{\Delta t}$ and r.

Interestingly, obeying equation 2.7 is sufficient for $q_{\Delta t}$ to satisfy equation 2.6. Precisely, let $q_{\Delta t}(x \to y)$ is a function that is smooth on Δt with $q_0(x \to y) = \delta(x - y)$, we are to show that, if $q_{\Delta t}$ satisfies equation 2.7, then it will obey equation 2.6. To do so, we take derivative on equation 2.7 by $\Delta t'$ at the origin, resulting in

$$q^{(1)}_{\Delta t}(x \to z) = \int_{\mathcal{X}} \mathrm{d}y q_{\Delta t}(x \to y) r(y, z),$$

where $r(z, y) := q^{(1)}_0(z|y)$. Then, we are to Taylor expand both sides by Δt. On the right hand side, we have

$$q_{\Delta t}(x \to y) = \delta(x - y) + \sum_{n=1}^{+\infty} \frac{(\Delta t)^n}{n!} q^{(n)}_0(x \to y),$$

and on the left hand side,

$$q^{(1)}_{\Delta t}(x \to z) = \sum_{n=0}^{+\infty} \frac{(\Delta t)^n}{n!} q^{(n+1)}_0(x \to z).$$

So, we get the Taylor expansion on both sides. At $(\Delta t)^0$ order,

$$q^{(1)}_0(x \to z) = r(x, z),$$

which is just the definition of r. At $(\Delta t)^1$ order,

$$q^{(2)}_0(x \to z) = \int_{\mathcal{X}} \mathrm{d}y q^{(1)}_0(x \to y) r(y, z) = \int_{\mathcal{X}} \mathrm{d}y r(x, y) r(y, z).$$

Iteratively at $(\Delta t)^n$ order, we will find

$$q^{(n)}_0(x \to z) = \int_{\mathcal{X}} \mathrm{d}y_1 \cdots \int_{\mathcal{X}} \mathrm{d}y_{n-1}\, r(x, y_1) \cdots r(y_{n-1}, z)$$

again. And this implies equation 2.6. So, we conclude this paragraph as follow: *obeying equation 2.7 is the essential and sufficient condition for a function $q_{\Delta t}(x|y)$, which is smooth on Δt with $q_0(x|y) = \delta(x - y)$, to satisfy equation 2.6; additionally, if $q_{\Delta t}(x|y)$ is non-negative and normalized on x (namely, $\int_{\mathcal{X}} \mathrm{d}x\, q_{\Delta t}(x|y) = 1$), then $q_{\Delta t}$ is a transition density.*

2.4 Detailed Balance Provides Stationary Distribution

Let Π a stationary solution of master equation 2.4. Then, its density function π satisfies $\int_{\mathcal{X}} \mathrm{d}x \pi(x) r(x, y) = 0$. Since we have demanded that $\int_{\mathcal{X}} \mathrm{d}x r(y, x) = 0$, the stationary master equation can be re-written as

$$\int_{\mathcal{X}} \mathrm{d}x \left[\pi(x) r(x, y) - \pi(y) r(y, x)\right] = 0.$$

But, this condition is too weak to be used. A more useful condition, which is stronger than this, is that the integrand vanishes everywhere:

$$\pi(x) r(x, y) = \pi(y) r(y, x), \tag{2.8}$$

which is called the **detailed balance condition**.

Interestingly, for a transition rate r that satisfies detailed balance condition 2.8, the transition density $q_{\Delta t}$ generated by r using equation 2.6 satisfies a similar relation

$$\pi(x) q_{\Delta t}(x \to y) = \pi(y) q_{\Delta t}(y \to x). \tag{2.9}$$

To see this, consider the third line in equation 2.6. It contributes to $\pi(x) q_{\Delta t}(x \to z)$ by the main factor $\int \mathrm{d}y \pi(x) r(x, y)\, r(y, z)$. By inserting the detailed balance condition for $r(x, y)$ as $\pi(x) r(x, y) = \pi(y) r(y, x)$, the main factor becomes $\int \mathrm{d}y r(y, x) \pi(y) r(y, z)$. Then inserting the detailed balance condition again, for $r(y, z)$ as $\pi(y) r(y, z) = \pi(z) r(z, y)$, results in $\int \mathrm{d}y \pi(z) r(z, y) r(y, x)$. It is the corresponding main factor that contributes to $\pi(z) q_{\Delta t}(z \to x)$. Applying the same process to other lines in equation 2.6, we arrive at equation 2.9.

2.5 Detailed Balance with Connectivity Monotonically Reduces Relative Entropy

Given the time t, if the time-dependent distribution $P(t)$ and the stationary distribution Π share the same alphabet $\mathcal{X}$, which means $p(x,t)>0$ and $\pi(x)>0$ for each $x\in\mathcal{X}$, then the relative entropy between them is well-defined, as

$$H(P(t),\Pi)=\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)\ln\frac{p(x,t)}{\pi(x)}. \tag{2.10}$$

It characterizes the difference between distributions $P(t)$ and Π. It is a plausible generalization of Shannon entropy to continuous random variables (see chapter 1).

We can calculate the time-derivative of relative entropy by master equation 2.4. Generally, the time-derivative of relative entropy has no interesting property. But, if the π is more than stationary but satisfying a stronger condition: the detailed balance condition, then $\mathrm{d}H(P(t),\Pi)/\mathrm{d}t$ will have a regular form[2.2]

$$\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)=-\frac{1}{2}\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\left(\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right). \tag{2.11}$$

2.2. The proof is given as follow. Directly, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)&=\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{X}}\mathrm{d}x\,[p(x,t)\ln p(x,t)-p(x,t)\ln\pi(x)]\\&=\int_{\mathcal{X}}\mathrm{d}x\left(\frac{\partial p}{\partial t}(x,t)\ln p(x,t)+\frac{\partial p}{\partial t}(x,t)-\frac{\partial p}{\partial t}(x,t)\ \ln\pi(x)\right).\end{aligned}$$

Since $\int_{\mathcal{X}}\mathrm{d}x\,(\partial p/\partial t)(x,t)=(\partial/\partial t)\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)=0$, the second term vanishes. Then, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)=\int_{\mathcal{X}}\mathrm{d}x\,\frac{\partial p}{\partial t}(x,t)\ln\frac{p(x,t)}{\pi(x)}.$$

Now, we replace $\partial p/\partial t$ by master equation 2.4, as

$$\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\,[p(y,t)r(y,x)-p(x,t)r(x,y)]\ln\frac{p(x,t)}{\pi(x)},$$

Then, insert detailed balance condition $r(y,x)=\pi(x)r(x,y)/\pi(y)$, as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)&=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\left(\frac{p(y,t)}{\pi(y)}\pi(x)r(x,y)-p(x,t)r(x,y)\right)\ln\frac{p(x,t)}{\pi(x)}\\&=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(y,t)}{\pi(y)}-\frac{p(x,t)}{\pi(x)}\right)\ln\frac{p(x,t)}{\pi(x)}.\end{aligned}$$

Since x and y are dummy, we interchange them in the integrand, and then insert detailed balance condition again, as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)&=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(y)r(y,x)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\ln\frac{p(y,t)}{\pi(y)}\\&=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\ln\frac{p(y,t)}{\pi(y)}.\end{aligned}$$

By adding the two previous results together, we find

$$\begin{aligned}&2\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)\\[\text{1st result}]&=-\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\ln\frac{p(x,t)}{\pi(x)}\\[\text{2nd result}]&+\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\ln\frac{p(y,t)}{\pi(y)}\\&=-\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\pi(x)r(x,y)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\left(\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right),\end{aligned}$$

from which we directly get the result. Notice that this proof is very tricky: it uses detailed balance condition twice, between which the expression is symmetrized. It is an ingenious mathematical engineering.

We are to check the sign of the integrand. The $r(x,y)$ is negative only when $x=y$, on which the integrand vanishes. Thus, $r(x,y)$ can be treated as non-negative, so is the $\pi(x)r(x,y)$ factor (since $\pi(x)>0$ for all $x\in\mathcal{X}$). Now, we check the sign of the last two terms. If $p(x,t)/\pi(x)>p(y,t)/\pi(y)$, then $\ln[p(x,t)/\pi(x)]>\ln[p(y,t)/\pi(y)]$, thus the sign of the last two terms is positive. The same goes for $p(x,t)/\pi(x)<p(y,t)/\pi(y)$. Only when $p(x,t)/\pi(x)=p(y,t)/\pi(y)$ can it be zero. Altogether, the integrand is non-positive, thus $\mathrm{d}H/\mathrm{d}t\leqslant 0$.

The integrand vanishes when either $r(x,y)=0$ or $p(x,t)/\pi(x)=p(y,t)/\pi(y)$. If $r(x,y)>0$ for each $x\neq y$, then $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ only when $p(x,t)/\pi(x)=p(y,t)/\pi(y)$ for all $x,y\in\mathcal{X}$, which implies that $p(\cdot,t)=\pi$ (since $\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)=\int_{\mathcal{X}}\mathrm{d}x\,\pi(x)=1$), or $P(t)=\Pi$.

Contrarily, if $r(x,y)=0$ on some subset $U\subset\mathcal{X}\times\mathcal{X}$, it seems that $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ cannot imply $p(x,t)/\pi(x)=p(y,t)/\pi(y)$ on U. But, if there is a $z\in\mathcal{X}$ such that both (x,z) and (y,z) are not in U, then $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ implies $p(x,t)/\pi(x)=p(z,t)/\pi(z)$ and $p(y,t)/\pi(y)=\pi(z,t)/\pi(z)$, thus implies $p(x,t)/\pi(x)=p(y,t)/\pi(y)$. It hints for connectivity. Precisely, for each $x,z\in\mathcal{X}$, if there is a series $(y_1,\ldots,y_n)$ from x $(y_1:=x)$ to z $(y_n:=z)$ with both $r(y_i,y_{i+1})$ and $r(y_{i+1},y_i)$ are positive for each i, then we say x and z are **connected**, and the series is called a **path**. It means *there are densities transiting along the forward and backward directions of the path*. In this situation, $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ implies $p(x,t)/\pi(x)=p(z,t)/\pi(z)$. We have, along the path, $p(y_1,t)/\pi(y_1)=p(y_2,t)/\pi(y_2)=\cdots=p(y_n,t)/\pi(y_n)$, thus $p(x,t)/\pi(x)=p(z,t)/\pi(z)$ since $x=y_1$ and $z=y_n$. So, by repeating the previous discussion in the case "$r(x,y)>0$ for each $x\neq y$", we find $P(t)=\Pi$ at $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ if every two elements in $\mathcal{X}$ are connected.

Let us examine the connectivity further. We additionally *define* that every element in $\mathcal{X}$ is connected to itself, then connectivity forms an equivalence relation. So, it separates $\mathcal{X}$ into subsets (equivalence classes) $\mathcal{X}_1,\ldots,\mathcal{X}_n$ with $\mathcal{X}_i\cap\mathcal{X}_j=\varnothing$ for each $i\neq j$ and $\mathcal{X}=\cup_{i=1}^n\mathcal{X}_i$. In each subset $\mathcal{X}_i$, every two elements are connected. In this way, the whole random system are separated into many independent subsystems. The distributions $P_i(t)$ and Π_i defined in the subsystem i have the alphabet $\mathcal{X}_i$ and density functions $p_i(x,t):=p(x,t)/\int_{\mathcal{X}_i}\mathrm{d}x\,p(x,t)$ and $\pi_i(x):=\pi(x)/\int_{\mathcal{X}_i}\mathrm{d}x\,\pi(x)$ respectively (the denominators are used for normalization). Applying the previous discussion to this subsystem, we find $P_i(t)=\Pi_i$ at $(\mathrm{d}/\mathrm{d}t)\,H(P_i(t),\Pi_i)=0$.

So, for the whole random system or each of its subsystems, the following theorem holds.

Theorem 2.1. *Let Π a distribution with alphabet $\mathcal{X}$. If there is a transition rate r such that 1) every two elements in $\mathcal{X}$ are connected and that 2) the detailed balance condition 2.8 holds for Π and r, then for any time-dependent distribution $P(t)$ with the same alphabet (at one time) evolved by the master equation 2.4, $P(t)$ will monotonically and constantly relax to Π.*

Many textures use Fokker-Planck equation to prove the monotonic reduction of relative entropy. After an integration by parts, they arrive at a negative definite expression, which means the monotonic reduction. This proof needs smooth structure on X, which is essential for integration by parts. In this section, we provides a more generic alternative to the proof, for which smooth structure on X is unnecessary.

2.6 Monte-Carlo Simulation and Guarantee of Relaxation

How to numerically simulate the evolution of master equation 2.4 that tends to equilibrium (at which the simulation terminates)? Using the simile of sands (see section 2.2), we simulate each sand, but replace its free will by a transition probability. Explicitly, we initialize the position of each sand randomly. Then iteratively update the positions. In each iteration, a sand jumps from position x to position y with the density function $q_{\Delta t}(x\to y)\approx\delta(x-y)+r(x,y)\,\Delta t$ where Δt is

sufficiently small. Not every transition is valid. On one hand, we have to ensure that computer has a sampler that makes random sampling for $q_{\Delta t}(x \to y)$. On the other hand, to ensure the termination, the transition rate r, together with the density function π, shall satisfy the detailed balance condition 2.8. Section 2.7 provides an algorithm that constructs such a transition rate from the density function $q_{\Delta t}$. Then, we *expect* that the simulation will iteratively decrease the difference between the distribution of the sands and the Π. We terminate the iteration when they have been close enough. In this way, we simulate a collection of sands evolves with the master equation to equilibrium, and finally distributes as Π. This process is called **Monte-Carlo simulation**, first developed by Stanislaw Ulam in 1940s while he was working on the project of nuclear weapons at Los Alamos National Laboratory. The name is in memory of Ulam's uncle who lost all his personal assets in Monte Carlo Casino, Monaco.[2.3]

Like the Euler method in solving dynamical system, however, a finite time step results in residual error. This residual error must be analyzed an controlled, so that the distribution will evolve toward Π, as we have expected. To examine this, we calculate the $H(P(t+\Delta t),\Pi) - H(P(t),\Pi)$ where Δt is small but still finite, and check when it is negative (such that $H(P(t))$ monotonically decreases to $P(t) \to \Pi$).

By definition, we have

$$\Delta H := H(P(t+\Delta t),\Pi) - H(P(t),\Pi) = \int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln \frac{p(x,t+\Delta t)}{\pi(x)} - \int_{\mathcal{X}} \mathrm{d}x\, p(x,t) \ln \frac{p(x,t)}{\pi(x)}.$$

Inserting $\int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln(p(x,t)/\pi(x,t))$ gives

$$\begin{aligned}
\Delta H &= \int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln \frac{p(x,t+\Delta t)}{\pi(x)} - \int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln \frac{p(x,t)}{\pi(x)} \\
&\quad + \int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln \frac{p(x,t)}{\pi(x)} - \int_{\mathcal{X}} \mathrm{d}x\, p(x,t) \ln \frac{p(x,t)}{\pi(x)} \\
&= \int_{\mathcal{X}} \mathrm{d}x\, p(x,t+\Delta t) \ln \frac{p(x,t+\Delta t)}{p(x,t)} \\
&\quad + \int_{\mathcal{X}} \mathrm{d}x\, [p(x,t+\Delta t) - p(x,t)] \ln \frac{p(x,t)}{\pi(x)}
\end{aligned}$$

The first line is recognized as $H(P(t+\Delta t), P(t))$, which is non-negative. Following the same steps in section 2.5 (but using discrete time master equation 2.5 instead, and detailed balance condition 2.9 for transition density), the second line reduces to

$$-\frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}y\, \pi(x) q_{\Delta t}(x \to y) \left(\frac{p(x,t)}{\pi(x)} - \frac{p(y,t)}{\pi(y)} \right) \left(\ln \frac{p(x,t)}{\pi(x)} - \ln \frac{p(y,t)}{\pi(y)} \right),$$

which is non-positive (suppose that r connects every two elements in $\mathcal{X}$). So, the sign of ΔH is determined by that which line has greater absolute value. The first line depends only on the difference between $P(t)$ and $P(t+\Delta t)$, thus Δt, while the second line additionally depends on the difference between $P(t)$ and Π (the factor $q_{\Delta t}(x|y)$ also depends on Δt). When $\Delta t \to 0$, the first line vanishes, while the second does not until $P(t) \to \Pi$. This suggests us to investigate how fast each term converges as $\Delta t \to 0$.

2.3. There are multiple motivations for Monte-Carlo simulation. An important one comes from numerical integration. The problem is calculating the integral $\int_{\mathcal{X}} \mathrm{d}x \pi(x)\, f(x)$ for a density function π and an arbitrary function $f: \mathcal{X} \to \mathbb{R}$. When $\mathcal{X}$ has finite elements, this integral is easy to compute, which is $\sum_{x \in \mathcal{X}} \pi(x)\, f(x)$. Otherwise, this integral will be intractable. Numerically, this integral becomes the expectation $(1/|\mathcal{S}|) \sum_{x \in \mathcal{S}} f(x)$ where $\mathcal{S}$ is a collection of elements randomly sampled from distribution Π (its density function is the π). By central limit theorem, the numerical error $|\int_{\mathcal{X}} \mathrm{d}x \pi(x)\, f(x) - (1/|\mathcal{S}|) \sum_{x \in \mathcal{S}} f(x)|$ is proportional to $1/\sqrt{|\mathcal{S}|}$, which can be properly bounded as long as $|\mathcal{S}|$ is large enough. But, how to sample from a distribution if you only know its density function (recall in section 1.2, a distribution is the combination of its density function and its sampler)? The answer is using Monte-Carlo simulation.

To examine the speed of convergence, we calculate the leading order of Δt in each line. To make it clear, we denote the first line by ΔH_1 and the second line by ΔH_2. Taylor expanding ΔH_1 by Δt gives[2.4]

$$\Delta H_1 = \frac{\Delta t^2}{2}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left(\frac{\partial}{\partial t}\ln p(x,t)\right)^2 + o(\Delta t^2),$$

where, by master equation 2.4, $(\partial/\partial t)\ln p(x,t) = \int_{\mathcal{X}} \mathrm{d}w\, p(w,t)\, r(w,x)/p(x,t)$. For ΔH_2, we insert equation 2.11 after Taylor expanding $q_{\Delta t}$ by Δt, and obtain

$$\Delta H_2 = \Delta t \frac{\mathrm{d}}{\mathrm{d}t} H(P(t), \Pi) + o(\Delta t).$$

We find ΔH_1 converges with speed Δt^2 while ΔH_2 has speed Δt.

Thus, given $P(t) \neq \Pi$ (so that $\Delta H_2 \neq 0$, recall section 2.5), there must be a $\delta > 0$ such that for any $\Delta t < \delta$, we have $|\Delta H_1| < |\Delta H_2|$, in which case the $\Delta H = \Delta H_1 + \Delta H_2 < 0$ (recall that $\Delta H_1 \geqslant 0$ and $\Delta H_2 \leqslant 0$). The δ is bounded by

$$\delta \leqslant \left[-\frac{\mathrm{d}}{\mathrm{d}t} H(P(t), \Pi)\right] \Big/ \left[\frac{1}{2}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left(\frac{\partial}{\partial t}\ln p(x,t)\right)^2\right].$$

This bound is proportional to the difference between $P(t)$ and Π (namely, the first factor). When $P(t)$ has approached Π (that is, $P(t) \approx \Pi$ but not exactly equal), δ has to be extremely small. (This is a little like supervised machine learning where Δt acts as learning rate and $H(P(t), \Pi)$ as loss. In the early stage of training, the loss has a greater slope of decrease and we can safely employ a relatively larger learning rate to speed up the training. But, we have to tune the learning rate to be smaller and smaller during the training, in which the slope of loss is gradually decreasing. Otherwise, the loss will not decrease but keep fluctuating when it has been sufficiently small, since the learning rate now becomes relatively too big.)

2.4. The first line

$$\Delta H_1 := \int_{\mathcal{X}} \mathrm{d}x\, p(x, t+\Delta t) \ln \frac{p(x,t+\Delta t)}{p(x,t)}$$

To Taylor expand the right hand side by Δt, we expand $p(x, t+\Delta t)$ to $o(\Delta t^2)$, as

$$p(x,t+\Delta t) = p(x,t) + \Delta t \frac{\partial p}{\partial t}(x,t) + \frac{\Delta t^2}{2!}\frac{\partial^2 p}{\partial t^2}(x,t) + o(\Delta t^2),$$

and the same for $\ln p(x, t+\Delta t)$, as

$$\ln p(x,t+\Delta t) = \ln p(x,t) + \Delta t \frac{\partial}{\partial t}\ln p(x,t) + \frac{\Delta t^2}{2!}\frac{\partial^2}{\partial t^2}\ln p(x,t) + o(\Delta t^2).$$

Plugging in $(\mathrm{d}/\mathrm{d}x)\ln f(x) = f'(x)/f(x)$ and then $(\mathrm{d}^2/\mathrm{d}x^2)\ln f(x) = f''(x)/f(x) - (f'(x)/f(x))^2$, we find

$$\ln p(x,t+\Delta t) - \ln p(x,t) = \Delta t\left[\frac{\partial p}{\partial t}p(x,t)\, p^{-1}(x,t)\right] + \frac{\Delta t^2}{2}\left[\frac{\partial^2 p}{\partial t^2}(x,t)\, p^{-1}(x,t) - \left(\frac{\partial p}{\partial t}(x,t)\, p^{-1}\right)^2\right] + o(\Delta t^2).$$

So, the Δt order term in ΔH_1 is

$$\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left[\frac{\partial p}{\partial t}p(x,t)\, p^{-1}(x,t)\right] = \int_{\mathcal{X}} \mathrm{d}x \frac{\partial p}{\partial t}p(x,t) = \frac{\partial}{\partial t}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t) = 0,$$

where we used the normalization of p. The Δt^2 term in ΔH_1 is

$$\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\,\frac{1}{2}\left[\frac{\partial^2 p}{\partial t^2}(x,t)\, p^{-1}(x,t) - \left(\frac{\partial p}{\partial t}p(x,t)\, p^{-1}(x,t)\right)^2\right] + \frac{\partial p}{\partial t}(x,t)\, p^{-1}(x,t)\frac{\partial p}{\partial t}(x,t).$$

Using the normalization of p as before, it is reduced to

$$\frac{1}{2}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left(\frac{\partial p}{\partial t}p(x,t)\, p^{-1}(x,t)\right)^2 = \frac{1}{2}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left(\frac{\partial}{\partial t}\ln p(x,t)\right)^2.$$

Altogether, we arrive at

$$\Delta H_1 = \frac{\Delta t^2}{2}\int_{\mathcal{X}} \mathrm{d}x\, p(x,t)\left(\frac{\partial}{\partial t}\ln p(x,t)\right)^2 + o(\Delta t^2).$$

2.7 Example: Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm is a simple method that constructs transition rate for any given stationary distribution such that detailed balance condition holds. Explicitly, given the density function of a stationary distribution Π, and an auxiliary transition rate γ (ensuring that $\gamma(x,y)>0$ for each x and y in alphabet $\mathcal{X}$ with $x\neq y$) the transition rate r is constructed by

$$r(x,y)=\gamma(x,y)\min\left(1,\frac{\pi(y)\gamma(y,x)}{\pi(x)\gamma(x,y)}\right). \tag{2.12}$$

This transition rate connects every two elements in $\mathcal{X}$ (since $\gamma(y,x)>0$ for each $x\neq y$). In addition, together with π, it satisfies detailed balance condition 2.8. By equation 2.12, we have

$$\pi(x)r(x,y)=\pi(x)\gamma(x,y)\min\left(1,\frac{\pi(y)\gamma(y,x)}{\pi(x)\gamma(x,y)}\right)=\min\left(\pi(x)\gamma(x,y),\pi(y)\gamma(y,x)\right).$$

We extract the second variable out of min function, makes it

$$\min\left(\frac{\pi(x)\gamma(x,y)}{\pi(y)\gamma(y,x)},1\right)\pi(y)\gamma(y,x)=\pi(y)\gamma(y,x)\min\left(1,\frac{\pi(x)\gamma(x,y)}{\pi(y)\gamma(y,x)}\right).$$

The right hand side is recognized as $\pi(y)r(y,x)$, implying that detailed balance condition 2.8 holds for the r constructed by equation 2.12. Theorem 2.1 then states that, *evolved by the master equation 2.4, any initial distribution will finally relax to the stationary distribution* Π.

Metropolis-Hastings algorithm was first proposed by Nicholas Metropolis and others in 1953 in Los Alamos, and then improved by Canadian statistician Wilfred Hastings in 1970. This algorithm was first defined for transition density. Together with a positive auxiliary transition density g, the transition density is defined as

$$q(x\to y):=g(x\to y)\min\left(1,\frac{\pi(y)g(u\to x)}{\pi(x)g(x\to y)}\right), \tag{2.13}$$

where g is positive-definite on $\mathcal{X}$. Notice that, in equation 2.13 there is no extra time parameter like the $q_{\Delta t}(x\to y)$ in section 2.2. It can be seen as a fixed time interval, which can only be used for discrete time master equation.

This definition has an intuitive and practical explanation. The two factors can be seen as two conditional probability. The factor $g(x\to y)$ first proposes a transition from x to y. (In numerical simulation, we have to ensure that computer has a sampler for sampling an x from the conditional density function $g(x\to y)$.) Then, this proposal will be accepted by Bernoulli probability with the ratio given by the first factor in the right hand side. If accepted, then transit to y, otherwise stay on x. Altogether, we get a conditional probability jumping from x to y, the $q(x\to y)$.

It is straight forward to check that, if, in addition, g smoothly depends on a parameter Δt as $g_{\Delta t}$, so is q as $q_{\Delta t}$, and if we Taylor expand $g_{\Delta t}$ at $\Delta t\to 0$ as $g_{\Delta t}(x\to y)=\delta(x-y)+\gamma(x,y)\,\Delta t+o(\Delta t)$, then we will find $q_{\Delta t}(x\to y)=\delta(x-y)+r(x,y)\,\Delta t+o(\Delta t)$. Indeed, when $x=y$, we have $q_{\Delta t}(x\to x)=g_{\Delta t}(x\to x)$. And when $x\neq y$, $\delta(x-y)=0$, we find

$$q_{\Delta t}(x\to y)=\left[\gamma(x,y)\min\left(1,\frac{\pi(y)\gamma(y,x)}{\pi(x)\gamma(x,y)}\right)\right]\Delta t+o(\Delta t).$$

Altogether, for each $x,y\in\mathcal{X}$, we find $q_{\Delta t}(x\to y)=\delta(x-y)+r(x,y)\,\Delta t+o(\Delta t)$. *In practice, we use the Metropolis-Hastings algorithm 2.13 to numerically simulate master equation 2.4.* But, based on the discussion in section 2.6, the Δt in $g_{\Delta t}$ shall be properly bounded to be small so as to ensure the relaxation $P(t)\to\Pi$.

Chapter 3

Kramers-Moyal Expansion and Langevin Process

We follow the discussion in chapter 2, but focusing on the specific situation where there is extra smooth structure on X. This smoothness reflects on the connectivity of the alphabet $\mathcal{X}$, and on the smooth "spatial" dependence of the density function and transition rate. This indicates that the conclusions in chapter 2 hold in this section, but the inverse is not guaranteed.

3.1 Conventions in This Chapter

Follow the conventions in chapter 2. In addition, thought this chapter, we set the alphabet to be Euclidean to get sufficient connectivity. Namely, $\mathcal{X} = \mathbb{R}^d$ for some integer $d \geqslant 1$.

We employ the **Einstein's convention**. That is, we omit the sum notation for the duplicated indices as long as they are "balanced". For example, $x_\alpha\, y^\alpha$ represents $\sum_\alpha x_\alpha\, y^\alpha$. The α appears twice in the expression, once in subscript (the x_α) and once in superscript (the y^α), for which we say indices are balanced. Expression like $x_\alpha\, y_\alpha$, however, does not represent a summation over α, because indices are not balanced (both are subscript). A more complicated example is $\partial_\alpha A^\alpha_\beta\, x^\beta$, which means $\sum_\alpha \sum_\beta \partial_\alpha A^\alpha_\beta\, x^\beta$. Einstein's convention is extremely helpful in taking Taylor expansion.

3.2 Cut-off in the Moments of Transition Rate Is Essential for Spatial Smoothness

For spatial smoothness, we assume that the density function $p(x,t)$ of a time-dependent distribution $P(t)$ are smooth on x and y. Besides the smoothness of functions, however, spatial smoothness demands more. To declare this, we consider the situation where the mass of $P(t)$ is centered at x initially, namely $p(x',0) = \delta(x-x')$. Then, after a small temporal period Δt, there is some portion of mass transits to elsewhere. By master equation 2.4, the change in density is

$$p(y,\Delta t) - p(y,0) = \Delta t \int_{\mathbb{R}^d} \mathrm{d}x'\, p(x',0) r(x',y) + o(\Delta t).$$

Inserting the initial condition $p(x',0) = \delta(x-x')$ and denoting $\epsilon := y - x$, we get

$$p(x+\epsilon,\Delta t) = \delta(\epsilon) + r(x,x+\epsilon)\,\Delta t + o(\Delta t).$$

We assume that the portion of mass does not transit far away from x, but in its neighbor, namely ϵ is really small in scale. Quantitatively, the scale is reflected by moments, where the nth-moment is defined by

$$M_n^{\alpha_1\cdots\alpha_n}(x,\Delta t) := \int_{\mathbb{R}^d} \mathrm{d}\epsilon p(x+\epsilon,\Delta t)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}). \tag{3.1}$$

Thus, the assumption turns to be

1. $M_n^{\alpha_1\cdots\alpha_n}(x,0) = 0$, and
2. as Δt tends to zero, $M_{n+m}(x,\Delta t)$ converges (to zero) faster than $M_{n+m}(x,\Delta t)$.

For the second condition, we shall expect that $(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n})$ will become much smaller when multiplied more small (random) variables.

Plugging in the expression of $p(x+\epsilon,\Delta t)$, we find

$$\begin{aligned} M_n^{\alpha_1\cdots\alpha_n}(x,\Delta t) &= \int_{\mathbb{R}^d} \mathrm{d}\epsilon\,\delta(\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}) + \Delta t\int_{\mathbb{R}^d}\mathrm{d}\epsilon\, r(x,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}) + o(\Delta t) \\ &= \Delta t\int_{\mathbb{R}^d}\mathrm{d}\epsilon\, r(x,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}) + o(\Delta t). \end{aligned}$$

Then, introducing (for distinguishing from moments, which is defined by density function, we using K instead of M for denoting the "moments of transition rate")

$$K_n^{\alpha_1\cdots\alpha_n}(x) := \int_{\mathbb{R}^d}\mathrm{d}\epsilon\, r(x,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}),$$

we have

$$M_n(x,\Delta t) = K_n(x)\,\Delta t + o(\Delta t).$$

So the first condition simply implies

$$\sup_{x\in\mathbb{R}^d}|K_n(x)| < +\infty. \tag{3.2}$$

The second condition is non-trivial. We are to show that it indicates a cut-off. That is, there shall exist an positive integer N_{cut}, such that $K_n = 0$ for any $n > N_{\mathrm{cut}}$. And we will find the estimation

$$M_n(x,\Delta t) = \mathcal{O}(\Delta t^{\sharp(n/N_{\mathrm{cut}})}), \tag{3.3}$$

where $\sharp$ is the ceiling function (which rounds its argument to the nearest greater integer). To obtain equation 3.3, we have to demand that (which will be realized later)

$$\sup_{x\in\mathbb{R}^d}|\partial_{\alpha_1}\cdots\partial_{\alpha_m}K_n(x)| < +\infty \tag{3.4}$$

holds for any indices $(\alpha_1,\ldots,\alpha_m)$. Together with equation 3.2, we demand that *the moments of transition rate and its derivatives are uniformly bounded on* $\mathbb{R}^d$. We devote the rest of this section to prove equation 3.3.

As an example for exploration, we first cut-off at $N_{\mathrm{cut}} = 2$, namely $K_n = 0$ for any $n > 2$. We are to calculate $\mathbb{E}[\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}]$ up to $o(\Delta t^2)$. This demands the relation between transition rate and transition density (equation 2.6),

$$q_{\Delta t}(x\to x+\epsilon) = \delta(\epsilon) + (\Delta t)\, r(x,x+\epsilon) + \frac{(\Delta t)^2}{2!}\int_{\mathcal{X}}\mathrm{d}y r(x,y)r(y,x+\epsilon) + o(\Delta t^2).$$

Starting at $p(y,0) = \delta(x-y)$, master equation gives

$$\begin{aligned} M_n^{\alpha_1\cdots\alpha_n}(x,\Delta t) &= \int_{\mathbb{R}^d}\mathrm{d}\epsilon\, p(x+\epsilon,\Delta t)\,\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n} \\ &= \int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y\, q_{\Delta t}(y\to x+\epsilon)\,\delta(x-y)\,\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n} \\ &= \int_{\mathbb{R}^d}\mathrm{d}\epsilon\, q_{\Delta t}(x\to x+\epsilon)\,\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}. \end{aligned}$$

Inserting the expansion of $q_{\Delta t}$ makes

$$\begin{aligned} M_n^{\alpha_1\cdots\alpha_n}(x,\Delta t) = \Delta t&\int_{\mathbb{R}^d}\mathrm{d}\epsilon\, r(x,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}) \\ &+ \frac{(\Delta t)^2}{2!}\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,y)r(y,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}) + o(\Delta t^2). \end{aligned}$$

The first term is $\Delta t K_n^{\alpha_1\cdots\alpha_n}(x)$, so it is $\Delta t K_1(x)$ and $\Delta t K_2(x)$ for $n=1,2$ respectively, and vanishes otherwise. In the rest, we focus on the second term, denoting the integral as

$$F_n^{\alpha_1\cdots\alpha_n}(x) := \int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,y)r(y,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}),$$

thus the second term is $\Delta t^2/2 F_n^{\alpha_1\cdots\alpha_n}(x)$. First notice that the integral has identity

$$\int_{\mathbb{R}^d} \mathrm{d}y r(x,y) r(y,x+\epsilon) = \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon).$$

Thus

$$F_n^{\alpha_1\cdots\alpha_n}(x) = \int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon)\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n}).$$

We explore the calculation of F_n by evaluating $F_1(x)$. By inserting an $(\epsilon - y)$ term, we get

$$\begin{aligned} F_1^{\alpha_1}(x) = &\int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+y+(\epsilon-y))\,(\epsilon^{\alpha_1} - y^{\alpha_1}) \\ &+ \int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon)\, y^{\alpha_1}. \end{aligned}$$

While integrating over ϵ, the first line gives $\int \mathrm{d}y r(x,x+y) K_1^{\alpha_1}(x+y)$, and the second vanishes because of the property $\int \mathrm{d}y r(x,y) = 0$. So, we find

$$F_1^{\alpha_1}(x) = \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) K_1^{\alpha_1}(x+y).$$

Taylor expansion of K_1 at x gives

$$\begin{aligned} &\int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y) K_1^{\alpha_1}(x+y) = K_1^{\alpha_1}(x) \int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y) + \\ &+ \partial_{\beta_1} K_1^{\alpha_1}(x) \int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y)\, y^{\beta_1} + \frac{1}{2!} \partial_{\beta_1}\partial_{\beta_2} K_1^{\alpha_1}(x) \int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y)\, y^{\beta_1} y^{\beta_2} \\ &+ \frac{1}{3!} \partial_{\beta_1}\partial_{\beta_2}\partial_{\beta_3} K_1^{\alpha_1}(x) \int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y)\, y^{\beta_1} y^{\beta_2} y^{\beta_3} + \cdots. \end{aligned}$$

The first term on the right hand side vanishes. The second term is $K_1^{\beta_1}(x)\,\partial_{\beta_1} K_1^{\alpha_1}(x)$, and the third is $(1/2!)\, K_2^{\beta_1\beta_2}(x)\,\partial_{\beta_1}\partial_{\beta_2} K_1^{\alpha_1}(x)$. The rest terms are all vanishing because $K_n = 0$ for any $n > 2$. So, we find

$$F_1^{\alpha_1}(x) = K_1^{\beta_1}(x)\,\partial_{\beta_1} K_1^{\alpha_1}(x) + \frac{1}{2} K_2^{\beta_1\beta_2}(x)\,\partial_{\beta_1}\partial_{\beta_2} K_1^{\alpha_1}(x). \tag{3.5}$$

Following the same process, we can obtain for $F_2(x)$,[3.1]

$$\begin{aligned} F_2^{\alpha_1\alpha_2}(x) = &K_1^{\beta_1}(x)\,\partial_{\beta_1} K_2^{\alpha_1\alpha_2}(x) + \frac{1}{2} K_2^{\beta_1\beta_2}(x) \partial_{\beta_1}\partial_{\beta_2} K_2^{\alpha_1\alpha_2}(x) + K_1^{\alpha_1}(x) K_1^{\alpha_2}(x) \\ &+ K_2^{\alpha_1\beta_1}(x) \partial_{\beta_1} K_1^{\alpha_2}(x) + \mathrm{perm}(\alpha_1,\alpha_2), \end{aligned}$$

3.1. Since

$$\epsilon^\alpha \epsilon^\beta = (\epsilon^\alpha - y^\alpha)(\epsilon^\beta - y^\beta) + (\epsilon^\alpha - y^\alpha) y^\beta + y^\alpha y^\beta + \mathrm{perm},$$

we have

$$\begin{aligned} F_2^{\alpha_1\alpha_2}(x) = &\int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon)\,(\epsilon^\alpha - y^\alpha)(\epsilon^\beta - y^\beta) \\ &+ \int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon)\,(\epsilon^\alpha - y^\alpha) y^\beta + \mathrm{perm} \\ &+ \int_{\mathbb{R}^d} \mathrm{d}\epsilon \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) r(x+y,x+\epsilon)\, y^\alpha y^\beta. \end{aligned}$$

Again, by integrating over ϵ, the first line on the right hand side is $\int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y) K_2^{\alpha_1\alpha_2}(x+y)$ and the last line vanishes. The second line is $\int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) K_1^{\alpha}(x+y) y^\beta + \mathrm{perm}$. Thus,

$$F_2^{\alpha_1\alpha_2}(x) = \int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y) K_2^{\alpha_1\alpha_2}(x+y) + \int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) K_1^{\alpha_1}(x+y) y^{\alpha_2} + \mathrm{perm}.$$

Then, again, Taylor expansion of Ks at x gives

$$\int_{\mathbb{R}^d} \mathrm{d}y\, r(x,x+y) K_2^{\alpha_1\alpha_2}(x+y) = K_1^{\beta_1}(x)\,\partial_{\beta_1} K_2^{\alpha_1\alpha_2}(x) + \frac{1}{2} K_2^{\beta_1\beta_2}(x) \partial_{\beta_1}\partial_{\beta_2} K_2^{\alpha_1\alpha_2}(x)$$

and

$$\int_{\mathbb{R}^d} \mathrm{d}y r(x,x+y) K_1^{\alpha_1}(x+y) y^{\alpha_2} = K_1^{\alpha_1}(x) K_1^{\alpha_2}(x) + K_2^{\alpha_2\beta_1}(x) \partial_{\beta_1} K_1^{\alpha_1}(x),$$

where we have used $K_n = 0$ for any $n > 2$ to cut the Taylor series. So, we find

$$F_2^{\alpha_1\alpha_2}(x) = K_1^{\beta_1}(x)\,\partial_{\beta_1} K_2^{\alpha_1\alpha_2}(x) + \frac{1}{2} K_2^{\beta_1\beta_2}(x) \partial_{\beta_1}\partial_{\beta_2} K_2^{\alpha_1\alpha_2}(x) + K_1^{\alpha_1}(x) K_1^{\alpha_2}(x) + K_2^{\alpha_1\beta_1}(x) \partial_{\beta_1} K_1^{\alpha_2}(x) + \mathrm{perm}.$$

where $\mathrm{perm}(\alpha_1,\alpha_2)$ permutes the α_1 and α_2 for any term that is not symmetric (which is the forth term on the right hand side). Then for $F_3(x)$, we have[3.2]

$$F_3^{\alpha_1\alpha_2\alpha_3}(x) = K_2^{\alpha_1\alpha_2}(x)K_1^{\alpha_3}(x) + K_2^{\alpha_1\beta_1}(x)\partial_{\beta_1}K_2^{\alpha_2\alpha_3}(x) + \mathrm{perm}(\alpha_1,\alpha_2,\alpha_3).$$

And for $F_4(x)$,

$$F_4^{\alpha_1\alpha_2\alpha_3\alpha_4}(x) = K_2^{\alpha_1\alpha_2}(x)\,K_2^{\alpha_3\alpha_4} + \mathrm{perm}(\alpha_1,\alpha_2,\alpha_3,\alpha_4).$$

When $n>4$, $F_n=0$ because $K_n=0$ for $n>2$. So, the F_n terminates at $n=4$.

For higher order expansion of $q_{\Delta t}$ by Δt (equation 2.6), the same goes. As an example, we examine the expansion at $\mathcal{O}(\Delta t^3)$, as

$$q_{\Delta t}(x\to x+\epsilon) = \cdots + \frac{\Delta t^3}{3!}\int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y'\, r(x,y)r(y,y'),r(y',x+\epsilon) + o(\Delta t^3),$$

where we have omitted the $\mathcal{O}(\Delta t^2)$ terms. Following the same derivation, we find it contributes to $\mathbb{E}[\epsilon^\alpha]$ by the term

$$G_1^\alpha(x) := \int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y'\, r(x,y)r(y,y'),r(y',x+\epsilon)\,\epsilon^\alpha.$$

We insert an $(\epsilon - y')$ term again, and get

$$\begin{aligned}G_1^\alpha(x) = &\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y' r(x,y)r(y,y'),r(y',x+\epsilon)\,(\epsilon^\alpha - y'^\alpha)\\ &+\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y'\, r(x,y)r(y,y'),r(y',x+\epsilon)\,y'^\alpha\end{aligned}$$

The second line vanishes after integrating over ϵ because $\int \mathrm{d}x\, r(x,y)=0$. The first line can be re-written as

$$\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y' r(x,y)r(y,x+y')r(x+y',(x+y')+(\epsilon-y'))\,(\epsilon^\alpha - y'^\alpha).$$

And integrating over ϵ gives

$$G_1^\alpha(x) = \int_{\mathbb{R}^d}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}y'\, r(x,y)r(y,x+y')K_1^\alpha(x+y').$$

3.2. Since

$$\begin{aligned}\epsilon^\alpha\epsilon^\beta\epsilon^\gamma = &(\epsilon^\alpha - y^\alpha)(\epsilon^\beta - y^\beta)(\epsilon^\gamma - y^\gamma)\\ &+(\epsilon^\alpha - y^\alpha)(\epsilon^\beta - y^\beta)y^\gamma + (\epsilon^\alpha - y^\alpha)(\epsilon^\gamma - y^\gamma)y^\beta + (\epsilon^\beta - y^\beta)(\epsilon^\gamma - y^\gamma)y^\alpha\\ &+(\epsilon^\alpha - y^\alpha)y^\beta y^\gamma + (\epsilon^\beta - y^\beta)y^\alpha y^\gamma + (\epsilon^\gamma - y^\gamma)y^\alpha y^\beta\\ &+y^\alpha y^\beta y^\gamma,\end{aligned}$$

we have

$$\begin{aligned}F_3^{\alpha_1\alpha_2\alpha_3}(x) = &\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)r(x+y,x+\epsilon)\,(\epsilon^{\alpha_1} - y^{\alpha_1})(\epsilon^{\alpha_2} - y^{a_2})(\epsilon^{\alpha_3} - y^{\alpha_3})\\ &+\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)r(x+y,x+\epsilon)\,(\epsilon^{\alpha_1} - y^{\alpha_1})(\epsilon^{\alpha_2} - y^{a_2})y^{\alpha_3} + \mathrm{perm}\\ &+\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)r(x+y,x+\epsilon)\,(\epsilon^{\alpha_1} - y^{\alpha_1})y^{\alpha_2}y^{\alpha_3} + \mathrm{perm}\\ &+\int_{\mathbb{R}^d}\mathrm{d}\epsilon\int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)r(x+y,x+\epsilon)\,y^{\alpha_1}y^{\alpha_2}y^{\alpha_3}.\end{aligned}$$

Again, using $K_n=0$ for any $n>2$, we get

$$F_3^{\alpha_1\alpha_2\alpha_3}(x) = \int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)K_2^{\alpha_1\alpha_2}(x+y)\,y^{\alpha_3} + \int_{\mathbb{R}^d}\mathrm{d}y r(x,x+y)K_1^{\alpha_1}(x+y)\,y^{\alpha_2}y^{\alpha_3} + \mathrm{perm}.$$

Taylor expansion of Ks at x gives

$$F_3^{\alpha_1\alpha_2\alpha_3}(x) = K_2^{\alpha_1\alpha_2}(x)K_1^{\alpha_3}(x) + K_2^{\alpha_1\beta_1}(x)\partial_{\beta_1}K_2^{\alpha_2\alpha_3}(x) + \mathrm{perm}.$$

Again, we Taylor expand K_1 at x, resulting in

$$\begin{aligned} G_1^\alpha(x) = K_1^\alpha(x) \int_{\mathbb{R}^d} \mathrm{d}y \int_{\mathbb{R}^d} \mathrm{d}y'\, r(x,y) r(y, x+y') \\ + \partial_{\beta_1} K_1^\alpha(x) \int_{\mathbb{R}^d} \mathrm{d}y \int_{\mathbb{R}^d} \mathrm{d}y'\, r(x,y) r(y, x+y')\, y'^{\beta_1} \\ + \frac{1}{2!} \partial_{\beta_1}\partial_{\beta_2} K_1^\alpha(x) \int_{\mathbb{R}^d} \mathrm{d}y \int_{\mathbb{R}^d} \mathrm{d}y'\, r(x,y) r(y, x+y')\, y'^{\beta_1} y'^{\beta_2} \\ + \cdots. \end{aligned}$$

By integrating over y, we find the first line vanishes because $\int \mathrm{d}x\, r(x,y) = 0$. We recognize that the second line is just the $F_1^\beta(x)$, and the third line is just the $F_2^{\beta_1\beta_2}(x)$. Since $F_n = 0$ for any $n > 4$, the series terminates at F_4. Thus, we arrive at

$$\begin{aligned} G_1^\alpha(x) = \; & F_1^{\beta_1}(x)\partial_{\beta_1} K_1^\alpha(x) + \frac{1}{2!} F_2^{\beta_1\beta_2}(x)\partial_{\beta_1}\partial_{\beta_2} K_1^\alpha(x) + \\ & \frac{1}{3!} F_3^{\beta_1\beta_2\beta_3}(x)\partial_{\beta_1}\partial_{\beta_2}\partial_{\beta_3} K_1^\alpha(x) + \frac{1}{4!} F_4^{\beta_1\beta_2\beta_3\beta_4}(x)\partial_{\beta_1}\partial_{\beta_2}\partial_{\beta_3}\partial_{\beta_4} K_1^\alpha(x). \end{aligned} \tag{3.6}$$

Comparing with F_1 (equation 3.5), we find the first line is the same as F_1 except that the K_ns (rather than the ∂K_ns) are replaced by F_n. Since F_n terminates at $n = 4$ (while K_n terminates at $n = 2$), there are more terms in G_1 than in F_1, namely the second line.

So, the problem for higher order expansion of $q_{\Delta t}$ is deduced to that we have solved at lower order. It means that we can calculate to arbitrary order of expansion iteratively, and the process is the same at each order.

Observing these results, we find the following rules (with explanations).

1. The superscripts are assigned to two Ks together with partial derivatives, ensuring that the extra indices (such as β_1) are all summed over (namely, contracted).

   → The reason why there are two Ks in F_n is that there are two rs in the expansion of $q_{\Delta t}$ at Δt^2 order.

2. For each nth order partial derivative, multiply it by a factor $1/n!$

   → Because partial derivative comes from Taylor expansion, and $1/n!$ is the coefficient.

3. Non-symmetric assignments have to be permuted.

   → For example, the $K_2^{\alpha_1\beta_1}(x)\partial_{\beta_1} K_1^{\alpha_2}(x)$ term in $F_2^{\alpha_1\alpha_2}(x)$ is not symmetric on α_1 and α_2, thus has to be permuted in the $\mathrm{perm}(\alpha_1, \alpha_2)$.

4. Symmetric assignments appear only once.

   → Also in $F_2^{\alpha_1\alpha_2}(x)$, the $K_1^{\alpha_1}(x) K_1^{\alpha_2}(x)$ term is symmetric on α_1 and α_2, so it appears only once, being absent in the $\mathrm{perm}(\alpha_1, \alpha_2)$.

5. We add all the possible assignments together as the final result.

These rules can be directly generalized to $\mathcal{O}(\Delta t^m)$ in the expansion of $q_{\Delta t}$ by r, in which there can be m Ks.

We have found that both $M_3(x, \Delta t)$ and $M_4(x, \Delta t)$ are of $\mathcal{O}(\Delta t^2)$, since both $F_3(x)$ and $F_4(x)$ are non-zero. But $F_5(x)$ must vanish since we cannot get five superscripts with only two K_ns with $n = 1, 2$. This further implies that any F_n with $n > 4$ is zero, leading $M_n(x, \Delta t)$ to $o(\Delta t^2)$. But at higher (than 2) order of Δt, there will be more (than two) Ks in $M_n(x, \Delta t)$. Then, based on the rules we just found, the number of combination of indices will be greater (than 4). These combinations, however, will always be "exhausted" when n has been sufficiently large. That is, there will be finite $M_n(x, \Delta t)$s at any given order of Δt. This is just the formal expression of the second condition we assumed. So, the assumption is guaranteed with cut-off. And conversely, only with a cut-off can we guarantee the assumption. This can be directly generalized to cut-off at any positive integer N_{cut}, namely $K_n = 0$ for any $n > N_{\mathrm{cut}}$. In this situation, we have (recall that $\sharp$ is the ceiling function) $M_n(x, \Delta t) = \mathcal{O}(\Delta t^{\sharp(n/N_{\mathrm{cut}})})$.

Besides, there is additional requirement on K_n. During the calculation, we have employed the condition that the partial derivatives of K_n are well-defined. That is, for any indices $(\alpha_1, \ldots, \alpha_m)$,

$$\sup_{x\in\mathbb{R}^d} |\partial_{\alpha_1}\cdots\partial_{\alpha_m} K_n(x)| < +\infty. \tag{3.7}$$

Together with equation 3.2, we conclude that *the moments of transition rate and its derivatives are uniformly bounded on* $\mathbb{R}^d$.

3.3 Transition Rate Is Determined by Its Moments

Since all Ks are finite, and we only have finite Ks, we can relate the Ks to the transition rate r explicitly. To do so, we first introduce an arbitrary test function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ in **Schwarts space** $S(\mathbb{R}^d)$, which is a functional space in which function mapping from $\mathbb{R}^d$ to $\mathbb{R}$ is smooth and rapidly tends to zero in the region far from origin.[3.3] For example, Gaussian function (the density function of normal distribution) is in Schwarts space $S(\mathbb{R})$. Since φ is in Schwarts space, in which functions are smooth, we can Taylor expanding φ at origin, which gives

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\,\varphi(\epsilon) = \int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\,\varphi(0) + \int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon) \sum_{n=1}^{+\infty} \frac{1}{n!} (\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n})\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0).$$

Because of the normalization of transition density, we have $\int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon) = 0$, thus

$$\begin{aligned}\int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\,\varphi(\epsilon) &= \int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon) \sum_{n=1}^{+\infty} \frac{1}{n!} (\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_n})\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0) \\ &= \sum_{n=1}^{N_{\mathrm{cut}}} \frac{1}{n!} K_n^{\alpha_1\cdots\alpha_n}(x)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0),\end{aligned}$$

where we have inserted the definition of Ks and that $K_n = 0$ for any $n > N_{\mathrm{cut}}$ (section 3.2). Then, because of the identity

$$(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0) = \int_{\mathbb{R}^d} \mathrm{d}\epsilon\,\delta(\epsilon)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(\epsilon),$$

integration by parts on the right hand side gives

$$(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0) = (-1)^n \int_{\mathbb{R}^d} \mathrm{d}\epsilon\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(\epsilon)\,\varphi(\epsilon).$$

Plugging this back, we find

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\,\varphi(\epsilon) = \int_{\mathbb{R}^d} \mathrm{d}\epsilon \left[\sum_{n=1}^{N_{\mathrm{cut}}} \frac{(-1)^n}{n!} K_n^{\alpha_1\cdots\alpha_n}(x)(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(\epsilon)\right]\varphi(\epsilon).$$

Since φ is arbitrary, we finall arrive at

$$r(x, x+\epsilon) = \sum_{n=1}^{N_{\mathrm{cut}}} \frac{(-1)^n}{n!} K_n^{\alpha_1\cdots\alpha_n}(x)(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(\epsilon). \tag{3.8}$$

This is called **Kramers–Moyal expansion**. It indicates that transition rate is determined by its moments K_ns. Transition rate has two continuous variables (namely, the x and y in $r(x, y)$), but each K_n has only one variable and there are only finite number of K_ns (section 3.2). So, *spatial smoothness greatly reduces the degree of freedom in stochastic process*.

3.3. Precisely, a function $\varphi \in S(\mathbb{R}^d)$ is infinitely differentiable (namely $\varphi \in C^\infty(\mathbb{R}^d)$) and satisfies the condition

$$\sup_{x\in\mathbb{R}^d} |x^{\alpha_1}\cdots x^{\alpha_m}(\partial_{\beta_1}\cdots\partial_{\beta_n}\varphi)(x)| < +\infty$$

for any indices $(\alpha_1, \ldots, \alpha_m)$ and $(\beta_1, \ldots, \beta_n)$. As $\|x\| \to +\infty$, φ (also its partial derivatives) falls faster than any polynomial (namely, the $x^{\alpha_1}\cdots x^{\alpha_m}$).

Because of the Dirac's delta functions, this transition rate is a generalized function. That is, only when applied to a test function can they be evaluated. For example, to evaluate $\partial_\alpha\delta(-x)$, we have to employ an arbitrary test function $\varphi\in S(\mathbb{R}^d)$, and calculate $\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi(x)$. First, notice that $\partial_\alpha\delta(-x)$ is in fact $(\partial_\alpha\delta)(-x)$ and that $(\partial\delta/\partial x^\alpha)(-x)=-(\partial/\partial x^\alpha)\delta(-x)$, thus

$$\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,(\partial_\alpha\delta)(-x)\,\varphi(x)=-\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha[\delta(-x)]\,\varphi(x).$$

Then, integration by parts gives $-\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha[\delta(-x)]\,\varphi(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\delta(-x)\,\partial_\alpha\varphi(x)$. After inserting the relation $\delta(x)=\delta(-x)$, we arrive at $\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi(x)=\partial_\alpha\varphi(0)$. On the other hand, we have, by integration by parts, $-\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(x)\,\varphi(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\delta(x)\,\partial_\alpha\varphi(x)=\partial_\alpha\varphi(0)$. Altogether, we find $\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi(x)=-\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\delta(x)\,\varphi(x)$, for any $\varphi\in S(\mathbb{R}^d)$. Thus, $\partial_\alpha\delta(-x)$ is evaluated to be $-\partial_\alpha\delta(x)$. That is, $\partial_\alpha\delta$ *is odd*. Following the same process, we can show that $\partial_\alpha\partial_\beta\delta$ *is even*.[3.4] These conclusions are to be used in section 3.9.

If we plug expansion 3.8 into master equation, then we get

$$\frac{\partial p}{\partial t}(x,t)=\int_{\mathbb{R}^d}\mathrm{d}w\,p(w,t)r(w,x)=\sum_{n=1}^{N_{\mathrm{cut}}}\frac{(-1)^n}{n!}\int_{\mathbb{R}^d}\mathrm{d}w\,p(w,t)K_n^{\alpha_1\cdots\alpha_n}(w)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x-w).$$

Notice that $(\partial/\partial w^{\alpha_1})\cdots(\partial/\partial w^{\alpha_n})\delta(x-w)=(-1)^n\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x-w)$, we get

$$\frac{\partial p}{\partial t}(x,t)=\sum_{n=1}^{N_{\mathrm{cut}}}\frac{1}{n!}\int_{\mathbb{R}^d}\mathrm{d}w\,p(w,t)K_n^{\alpha_1\cdots\alpha_n}(w)\left[\frac{\partial}{\partial w^{\alpha_1}}\cdots\frac{\partial}{\partial w^{\alpha_n}}\delta(x-y)\right].$$

If $p(\cdot,t)\in S(\mathbb{R}^d)$, then we have $p(x,t)\to 0$ as $\|x\|\to+\infty$. It means we can take integration by parts on the right hand side as (boundary terms vanish)

$$\sum_{n=1}^{N_{\mathrm{cut}}}\frac{(-1)^n}{n!}\int_{\mathbb{R}^d}\mathrm{d}w\,\delta(x-w)\left(\frac{\partial}{\partial w^{\alpha_1}}\cdots\frac{\partial}{\partial w^{\alpha_n}}\right)[K_n^{\alpha_1\cdots\alpha_n}(w)p(w,t)].$$

Then, integrating over w gives

$$\frac{\partial p}{\partial t}(x,t)=\sum_{n=1}^{N_{\mathrm{cut}}}\frac{(-1)^n}{n!}\left(\frac{\partial}{\partial x^{\alpha_1}}\cdots\frac{\partial}{\partial x^{\alpha_n}}\right)[K_n^{\alpha_1\cdots\alpha_n}(x)\,p(x,t)].\tag{3.9}$$

This is the form of Kramers–Moyal expansion that appears in many textures.

If $p(\cdot,t)\in S(\mathbb{R}^d)$, then $p(\cdot,t)$ as well as its partial derivatives rapidly tend to zero in the region far from origin. Together with condition 3.7 (K_n and its partial derivatives are bounded), we can show that $(\partial p/\partial t)(\cdot,t)\in S(\mathbb{R}^d)$ too. It then implies that $p(\cdot,t')\in S(\mathbb{R}^d)$ for any $t'>t$. That is, *if a time-dependent density function $p(\cdot,t)$ sits in Schwartz space initially, then it stays in Schwartz space during the evolution.*

3.4. We are to calculate $\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(-x)\,f(x)$, where $f\in S(\mathbb{R}^d)$. Again, noticing that $(\partial_\alpha\partial_\beta\delta)(-x)=\partial_\alpha\partial_\beta[\delta(-x)]$, we have

$$\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(-x)\,f(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,(\partial_\alpha\partial_\beta\delta)(-x)\,f(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta[\delta(-x)]\,f(x).$$

Then integration by parts gives

$$\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta[\delta(-x)]\,f(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\delta(-x)\,\partial_\alpha\partial_\beta f(x)=\partial_\alpha\partial_\beta f(0).$$

That is, $\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(-x)\,f(x)=\partial_\alpha\partial_\beta f(0)$. On the other hand, we have

$$\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(x)\,f(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\delta(x)\,\partial_\alpha\partial_\beta f(x)=\partial_\alpha\partial_\beta f(0).$$

So,

$$\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(-x)\,f(x)=\int_{\mathbb{R}^d}\mathrm{d}x\,\partial_\alpha\partial_\beta\delta(x)\,f(x)$$

holds for any $f\in S(\mathbb{R}^d)$, thus $\partial_\alpha\partial_\beta\delta(-x)=\partial_\alpha\partial_\beta\delta(x)$.

3.4 Randomness Is Absent in the First Moment of Transition Rate

In section 3.2, we have found that a finite cut-off on the moments of transition rate is essential for spatial smoothness. We are to show that cut-off at $N_{\text{cut}} = 1$ (namely $K_n = 0$ for any $n > 1$) only results in a deterministic evolution. With $N_{\text{cut}} = 1$, equation 3.9 reduces to (re-denote K_1 to f for simplicity)

$$\frac{\partial p}{\partial t}(x,t) + \partial_\alpha(f^\alpha(x)\, p(x,t)) = 0.$$

This is the **continuity equation** or **transport equation**. It was used for describing the evolution of the density of incompressible liquids. We are to solve this equation explicitly. As a first order partial differential equation, we can use the **method of characteristics**. At the first step, we fully expand the equation, as

$$\frac{\partial p}{\partial t}(x,t) + f^\alpha(x)\, \partial_\alpha p(x,t) = -\partial_\alpha f^\alpha(x)\, p(x,t).$$

The next step is constructing a parameterized curve $(x(s), t(s))$ for $s \in [0, +\infty)$ called **characteristic**, obeying

$$\frac{\mathrm{d}t}{\mathrm{d}s}(s) = 1$$

and

$$\frac{\mathrm{d}x^\alpha}{\mathrm{d}s}(s) = f^\alpha(x(s)).$$

It has solution $t(s) = s + t(0)$. If we set $t(0) = 0$, then we have $t = s$ and

$$\frac{\mathrm{d}x^\alpha}{\mathrm{d}s}(s) = \frac{\mathrm{d}x^\alpha}{\mathrm{d}t}(t) = f^\alpha(x(t)),$$

from which we solve $x(t)$, leading to

$$\frac{\mathrm{d}}{\mathrm{d}t}p(x(t),t) = \frac{\partial p}{\partial x^\alpha}(x(t),t)\,\frac{\mathrm{d}x^\alpha}{\mathrm{d}t}(t) + \frac{\partial p}{\partial t}(x(t),t) = \partial_\alpha p(x(t),t)\; f^\alpha(x(t)) + \frac{\partial p}{\partial t}(x(t),t),$$

which is the left hand side of the original equation along characteristic. Thus, by equaling to the right hand side of the original equation, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}p(x(t),t) = -\partial_\alpha f^\alpha(x(t))\, p(x(t),t).$$

Now, the original partial differential equation is converted into an ordinary differential equation. It has the unique solution

$$p(x(t),t) = p(x(0),0) \times \exp\left(-\int_0^t \mathrm{d}t'\, \partial_\alpha f^\alpha(x(t')) \right).$$

It indicates that the density at $x(0)$ will "transport" along the curve $x(t)$ as time evolves. For example, consider $p(y,0) = \delta(y - x(0))$, that is all mass is centered at $x(0)$. Then $p(x,t)$ will have density only at $x(t)$, and $p(y,t) = 0$ for any $y \neq x(t)$, since y is traced back to $y(0)$ along the characteristic $y(t)$ and $y(0) \neq x(0)$. That is to say, transport equation is deterministic.

3.5 Randomness Appears in the Second Moment of Transition Rate

In section 3.4, we have analyzed the cut-off at $N_{\text{cut}} = 1$ and found it deterministic, thus not a stochastic process. It indicates that we have to cut-off at least at $N_{\text{cut}} = 2$. We are to show that, if K_2 as a matrix-valued field is positive definite, then the randomness of Markovian process is guaranteed.

We examine this by using an example. Let K_1 is everywhere vanishing and K_2 is a constant identity matrix. Then, equation 3.9 reduces to

$$\frac{\partial p}{\partial t}(x,t) = \frac{1}{2}\delta^{\alpha\beta}\partial_\alpha\partial_\beta p(x,t). \tag{3.10}$$

This equation is the famous **heat equation** or **diffusion equation**, first investigated by French mathematician Joseph Fourier in 1822 for modeling how heat diffuses. His method is now named as **Fourier transformation**.

Define Fourier transformation of $p(x,t)$ on x as

$$\hat{p}(k,t) := \int_{\mathbb{R}^d} \mathrm{d}x \exp(-\mathrm{i}k_\alpha x^\alpha)\, p(x,t).$$

It has an inverse transformation

$$p(x,t) = \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp(\mathrm{i}k_\alpha x^\alpha)\, \hat{p}(k,t).$$

This relation holds because, by plugging $\hat{p}(k,t)$ into the right hand side, we find it

$$\int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp(\mathrm{i}k_\alpha x^\alpha) \int_{\mathbb{R}^d} \mathrm{d}y \exp(-\mathrm{i}k_\alpha y^\alpha)\, p(y,t) = \int_{\mathbb{R}^d} \mathrm{d}y \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp(\mathrm{i}k_\alpha (x^\alpha - y^\alpha))\, p(y,t).$$

Integrating over k and using the relation[3.5]

$$\delta(x-y) = \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp(\mathrm{i}k_\alpha (x^\alpha - y^\alpha)),$$

we find the right hand side $\int \mathrm{d}y \delta(x-y)\, p(y,t) = p(x,t)$ which is the left hand side. By plugging this inverse transformation into heat equation 3.10, we get

$$\int_{\mathbb{R}^d} \frac{\mathrm{d}k'}{(2\pi)^d} \exp(\mathrm{i}k'_\alpha x^\alpha) \left[\frac{\partial \hat{p}}{\partial t}(k',t) + \frac{1}{2}\delta^{\alpha\beta} k'_\alpha k'_\beta \hat{p}(k',t) \right] = 0.$$

Multiplied by $\exp(-\mathrm{i}k_\alpha x^\alpha)$ on both sides and integrating over x and then k', we arrive at

$$\frac{\partial \hat{p}}{\partial t}(k,t) + \frac{1}{2}\delta^{\alpha\beta} k_\alpha k_\beta \hat{p}(k,t) = 0.$$

This is an ordinary differential equation for each k. It has solution

$$\hat{p}(k,t) = \hat{p}(k,0) \exp\left(-\frac{1}{2}\delta^{\alpha\beta} k_\alpha k_\beta t \right).$$

Thus, by the inverse transformation

$$p(x,t) = \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d}\, \hat{p}(k,0) \exp\left(\mathrm{i}k_\alpha x^\alpha - \frac{1}{2}\delta^{\alpha\beta} k_\alpha k_\beta t \right).$$

Inserting the definition of $\hat{p}(k,0)$, we get

$$p(x,t) = \int_{\mathbb{R}^d} \mathrm{d}w\, p(w,0) \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp\left(\mathrm{i}k_\alpha (x^\alpha - w^\alpha) - \frac{1}{2}\delta^{\alpha\beta} k_\alpha k_\beta t \right).$$

The second integral is Gaussian. By the formula of Gaussian integral, which holds for any positive definite matrix A,

$$\int_{\mathbb{R}^d} \mathrm{d}x \exp\left(-\frac{1}{2} A_{\alpha\beta} x^\alpha x^\beta + b_\alpha x^\alpha \right) = \sqrt{\frac{(2\pi)^d}{\det A}} \exp\left(\frac{1}{2} (A^{-1})^{\alpha\beta} b_\alpha b_\beta \right),$$

we can integrate the second integral and find (replacing $A_{\alpha\beta}$ by $\delta^{\alpha\beta} t$ and b_α by $\mathrm{i}\,(x^\alpha - w^\alpha)$)

$$p(x,t) = \int_{\mathbb{R}^d} \mathrm{d}w\, p(w,0) \left[\frac{1}{\sqrt{(2\pi t)^d}} \exp\left(-\frac{1}{2t} \delta_{\alpha\beta} (x^\alpha - w^\alpha)(x^\beta - w^\beta) \right) \right].$$

From this expression, we can read out the transition density directly, as

$$q_t(w \to x) = \prod_{\alpha=1}^{d} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2t} (x^\alpha - w^\alpha)^2 \right), \tag{3.11}$$

3.5. TODO: explain this.

which obeys a normal distribution with w as its mean and t as its variance. So, we find randomness in the cut-off $N_{\text{cut}} = 2$.

Historically, in 1827, botanist Robert Brown noticed that pollen particles automatically shakes in water. It was first explained by Albert Einstein in 1905. He argued that the pollen particles are constantly stricken by water molecules, and found the transition density to be equation 3.11. Hence, the stochastic process described by this transition density is named by **Brownian motion**. Even though the techniques used for deriving this transition density had been mature when Brown first observed this phenomenon, but almost one hundred years after Brown's discover, in 1918, Norbert Wiener first constructed a complete mathematical theory for this stochastic process. So, it is also called **Wiener process**.

3.6 Langevin Process Is a Markovian Process with $N_{\text{cut}} = 2$

The most generic form of transition rate with $N_{\text{cut}} = 2$ is (by Kramers–Moyal expansion 3.8 with $N_{\text{cut}} = 2$, and re-denote K_1 by f and K_2 by Σ)

$$r(x, x+\epsilon) = -f^\alpha(x)\,\partial_\alpha\delta(\epsilon) + \frac{1}{2}\Sigma^{\alpha\beta}(x)\,\partial_\alpha\partial_\beta\delta(\epsilon), \tag{3.12}$$

A (continuous time) Markovian process with this transition rate is called a **Langevin process** or **Langevin dynamics**. It was first developed by French physicist Paul Langevin in 1908.

Plugging equation 3.12 into master equation 3.9, we find

$$\frac{\partial p}{\partial t}(x,t) = -\partial_\alpha(f^\alpha(x)\,p(x,t)) + \frac{1}{2}\partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,p(x,t)). \tag{3.13}$$

This equation is called **Fokker-Planck equation**, found by Adriaan Fokker and Max Planck in 1914 and 1917 respectively, or **Kolmogorov forward equation**, independently discovered in 1931.

As a matrix-valued field, Σ is symmetric and everywhere positive definite. Symmetry means $\Sigma^{\alpha\beta}(x) = \Sigma^{\beta\alpha}(x)$. This is a direct result of its definition $\int \mathrm{d}\epsilon r(x, x+\epsilon)\,\epsilon^\alpha\epsilon^\beta$. To see why it is positive definite, we consider the expectation

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon\, q_{\Delta t}(x \to x+\epsilon)\,\epsilon^\alpha\epsilon^\beta = \Sigma^{\alpha\beta}(x)\,\Delta t + o(\Delta t).$$

Under a proper coordinate of ϵ, it becomes diagonalized with the diagonal element (which is the eigenvalue)

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon\, q_{\Delta t}(x \to x+\epsilon)\,(\epsilon^\alpha)^2 > 0.$$

So, $\Sigma(x)$ is positive definite for any $x \in \mathbb{R}^d$.

3.7 Transition Density of Langevin Process Is Nearly Gaussian

The transition density of Langevin process is hard to obtain, except for some very simple situations (such as that in section 3.5). Even though, it can be properly approximated by Gaussian density function. Explicitly, let $q_{\Delta t}(x \to y)$ denote the transition density of a Langevin process. As in section 3.6, we re-denote K_1 by f and K_2 by Σ. Then, consider the Gaussian conditional density function (which may not be a transition density)

$$\tilde{q}_{\Delta t}(x \to x+\epsilon) := \frac{1}{\sqrt{(2\pi\Delta t)^d \det\Sigma(x)}} \exp\left(-\frac{1}{2\Delta t}[\Sigma^{-1}(x)]_{\alpha\beta}\,[\epsilon^\alpha - f^\alpha(x)\Delta t][\epsilon^\beta - f^\beta(x)\Delta t]\right).$$

We are to show that, for any function φ in Schwartz space $S(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon\, q_{\Delta t}(x \to x+\epsilon)\varphi(\epsilon) = \int_{\mathbb{R}^d} \mathrm{d}\epsilon\, \tilde{q}_{\Delta t}(x \to x+\epsilon)\varphi(\epsilon) + o(\Delta t). \tag{3.14}$$

That is, the transition density of Langevin process is approximated by a Gaussian conditional density function, *in the sense of applying onto Schwartz space*.

To prove equation 3.14, we first notice that the left hand side can be expanded by Δt as

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon q_{\Delta t}(x \to x+\epsilon)\varphi(\epsilon) = \varphi(0) + \Delta t \int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\varphi(\epsilon) + o(\Delta t).$$

So, all we need to do is proving that

$$\int_{\mathbb{R}^d} \mathrm{d}\epsilon \tilde{q}_{\Delta t}(x \to x+\epsilon)\varphi(\epsilon) = \varphi(0) + \Delta t \int_{\mathbb{R}^d} \mathrm{d}\epsilon r(x, x+\epsilon)\varphi(\epsilon) + o(\Delta t).$$

To show this, we Taylor expand φ at origin. The right hand side becomes

$$\varphi(0) + \Delta t \partial_\alpha \varphi(0) f^\alpha(x) + \frac{1}{2}\Delta t \partial_\alpha \partial_\beta \varphi(0) \Sigma^{\alpha\beta}(x) + o(\Delta t),$$

where we have inserted the moments of transition rate with cut-off $N_{\mathrm{cut}} = 2$. Since $\tilde{q}_{\Delta t}(x \to y)$ is the density function of the normal distribution with mean $x + f(x)\,\Delta t$ and covariance $\Sigma(x)\,\Delta t$, the left hand side is evaluated to be

$$\varphi(0) + \partial_\alpha \varphi(0) f^\alpha(x)\Delta t + \frac{1}{2}\partial_\alpha \partial_\beta \varphi(0)\Sigma^{\alpha\beta}(x)\Delta t + o(\Delta t),$$

which equals to the right hand side. Thus equation 3.14 holds.

Because of this Gaussian approximation, we can approximate Langevin process using a stochastic difference equation

$$X^\alpha(t+\Delta t) \approx X^\alpha(t) + f^\alpha(X)\,\Delta t + \eta^\alpha(X),$$

where $\eta(X)$ obeys a normal distribution with zero mean and covariance $\Sigma(X)\,\Delta t$ (this is why η depends on X). We can separate the dependence of X in η using **Cholesky factorization**.

To introduce Cholesky factorization, we fix the argument x and omit it for simplicity, so $\Sigma(x)$ is written as Σ. Since Σ is symmetric and positive definite (proved in section 3.6), we can diagonalize it using an orthogonal matrix E as $\Sigma = E^T \Lambda E$, where the diagonal $\Lambda_{\alpha\beta} = \delta_{\alpha\beta}\lambda_\beta$ with $\lambda_\beta > 0$. Define $\sqrt{\Lambda}_{\alpha\beta} := \delta_{\alpha\beta}\sqrt{\lambda_\beta}$, thus $\Lambda = \sqrt{\Lambda}^T \sqrt{\Lambda}$, and $\Sigma = C^T C$ where $C := \sqrt{\Lambda}\,E$. We thus factorize Σ into the "square" of C. Notice that C is invertible, and $C^{-1} = E^T (\sqrt{\Lambda})^{-1}$. This was first discovered by French military officer André-Louis Cholesky, who was killed in battle a few months before the end of World War I, dead at age 31. So, we have (insert the omitted x again) $\Sigma(x) = C^T(x)\,C(x)$ and the stochastic difference equation comes to be

$$X^\alpha(t+\Delta t) \approx X^\alpha(t) + f^\alpha(X(t))\,\Delta t + C^\alpha_\beta(X(t))\Delta W^\beta,$$

where ΔW^α obeys a standard normal distribution (we use W to indicate Wiener process).

In the limit $\Delta t \to 0$, the approximation becomes exact (so $\approx$ is replaced by $=$), and the equation turns from difference to be differential, as

$$\frac{\mathrm{d}X^\alpha}{\mathrm{d}t}(t) = f^\alpha(X(t)) + C^\alpha_\beta(X(t))\frac{\mathrm{d}W^\beta}{\mathrm{d}t}(t) \tag{3.15}$$

with

$$\mathbb{E}\left[\frac{\mathrm{d}W^\alpha}{\mathrm{d}t}(t)\frac{\mathrm{d}W^\beta}{\mathrm{d}t}(t')\right] = \delta^{\alpha\beta}\delta(t-t'). \tag{3.16}$$

This is the format that Langevin process appears in many textbooks.

3.8 Stationary Solution of Langevin Process Has Source-Free Degree of Freedom

In this section and section 3.9, we use Langevin process as an example to explicitly show what the detailed balance condition appends to the stationary condition.

The master equation of Langevin process (equation 3.13) has stationary solution Π which satisfies (since there is only one variable x, we use ∂ instead of ∇)

$$-\partial_\alpha(f^\alpha(x)\,\pi(x))+\frac{1}{2}\partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x))=0,$$

which means

$$f^\alpha(x)\,\pi(x)=\frac{1}{2}\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x))+\nu^\alpha(x), \tag{3.17}$$

where $\nu\colon\mathbb{R}^d\to\mathbb{R}^d$ is an arbitrary vector field such that $\partial_\alpha\nu^\alpha(x)=0$.

The vector field ν has an intuitive explanation. Regarding ν as a flux on $\mathbb{R}^d$, we find that there is not net flux flowing out of anywhere in $\mathbb{R}^d$. Otherwise, suppose there is $x\in\mathbb{R}^d$ and a closed surface S around x such that the net flux $\int \mathrm{d}S\cdot\nu(x)$ does not vanish. Then, by Stokes theorem, the surface integral $\int \mathrm{d}S\cdot\nu(x)=\int \mathrm{d}x\nabla\cdot v(x)=0$, thus conflicts. Such a vector field ν is called **free of source** or **source-free**.

3.9 Detailed Balance of Langevin Process Lacks Source-Free Degree of Freedom

After discussing stationary distribution of Fokker-Planck equation (as a master equation), we continue investigate when will Langevin process relax an initial distribution to the stationary. By theorem 2.1, this is equivalent to ask: when will the transition rate of Langevin process satisfy detailed balance condition? Detailed balance condition reads $r(x+\epsilon,x)\,\pi(x)=r(x,x+\epsilon)\,\pi(x+\epsilon)$. Directly inserting equation 3.12, we get for the left hand side,

$$\pi(x)r(x,x+\epsilon)=-\pi(x)f^\alpha(x)\,\partial_\alpha\delta(\epsilon)+\frac{1}{2}\pi(x)\Sigma^{\alpha\beta}(x)\,\partial_\alpha\partial_\beta\delta(\epsilon).$$

For the right hand side, we first have

$$\pi(x+\epsilon)\,r(x+\epsilon,x)=\pi(x+\epsilon)\,r(x+\epsilon,(x+\epsilon)-\epsilon).$$

Then, inserting equation 3.12 gives

$$\pi(x)r(x,x+\epsilon)=-\pi(x+\epsilon)f^\alpha(x+\epsilon)\partial_\alpha\delta(-\epsilon)+\frac{1}{2}\pi(x+\epsilon)\Sigma^{\alpha\beta}(x+\epsilon)\partial_\alpha\partial_\beta\delta(-\epsilon).$$

Since $\partial_\alpha\delta(-x)=-\partial_\alpha\delta(x)$ and $\partial_\alpha\partial_\beta\delta(-x)=\partial_\alpha\partial_\beta\delta(x)$ (derived in the end of section 3.3), it turns to be

$$\pi(x)r(x,x+\epsilon)=\pi(x+\epsilon)f^\alpha(x+\epsilon)\partial_\alpha\delta(\epsilon)+\frac{1}{2}\pi(x+\epsilon)\Sigma^{\alpha\beta}(x+\epsilon)\partial_\alpha\partial_\beta\delta(\epsilon).$$

As generalized functions, we are to examine these two expressions by using an arbitrary test function φ in Schwartz space $S(\mathbb{R}^d)$. Applying $\pi(x)r(x,x+\epsilon)$ to φ gives

$$\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x)r(x,x+\epsilon)\varphi(\epsilon)=-\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x)f^\alpha(x)\partial_\alpha\delta(\epsilon)\varphi(\epsilon)+\frac{1}{2}\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x)\Sigma^{\alpha\beta}(x)\partial_\alpha\partial_\beta\delta(\epsilon)\varphi(\epsilon).$$

Integration by parts gives (note that the ∂ is applied on ϵ)

$$\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x)r(x,x+\epsilon)\varphi(\epsilon)=\pi(x)\,f^\alpha(x)\partial_\alpha\varphi(0)+\frac{1}{2}\pi(x)\Sigma^{\alpha\beta}(x)\partial_\alpha\partial_\beta\varphi(0).$$

By applying $\pi(x+\epsilon)r(x+\epsilon,x)$ to φ, we get

$$\begin{aligned}\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x+\epsilon)r(x+\epsilon,x)\varphi(\epsilon)=&\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x+\epsilon)f^\alpha(x+\epsilon)\partial_\alpha\delta(\epsilon)\varphi(\epsilon)\\&+\frac{1}{2}\int_{\mathbb{R}^d}\mathrm{d}\epsilon\pi(x+\epsilon)\Sigma^{\alpha\beta}(x+\epsilon)\partial_\alpha\partial_\beta\delta(\epsilon)\varphi(\epsilon).\end{aligned}$$

Again, integration by parts results in (again, the ∂ operator is applied on ϵ)

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathrm{d}\epsilon\, \pi(x+\epsilon) r(x+\epsilon, x)\varphi(\epsilon) \\
=& -\int_{\mathbb{R}^d} \mathrm{d}\epsilon \delta(\epsilon) \frac{\partial}{\partial \epsilon^\alpha}[\pi(x+\epsilon) f^\alpha(x+\epsilon)\varphi(\epsilon)] + \frac{1}{2}\int_{\mathbb{R}^d} \mathrm{d}\epsilon \delta(\epsilon) \frac{\partial^2}{\partial\epsilon^\alpha \partial\epsilon^\beta}[\pi(x+\epsilon)\Sigma^{\alpha\beta}(x+\epsilon)\varphi(\epsilon)] \\
=& -\partial_\alpha[\pi(x) f^\alpha(x)]\, \varphi(0) - \pi(x) f^\alpha(x)\partial_\alpha\varphi(0) \\
&+ \frac{1}{2}\partial_\alpha\partial_\beta[\pi(x)\Sigma^{\alpha\beta}(x)]\, \varphi(0) + \partial_\beta[\pi(x)\Sigma^{\alpha\beta}(x)]\, \partial_\alpha\varphi(0) + \frac{1}{2}\pi(x)\Sigma^{\alpha\beta}(x)\partial_\alpha\partial_\beta\varphi(0).
\end{aligned}$$

By equaling $\int \mathrm{d}\epsilon\, \pi(x) r(x, x+\epsilon)\varphi(\epsilon)$ and $\int \mathrm{d}\epsilon\, \pi(x+\epsilon) r(x+\epsilon, x)\varphi(\epsilon)$, since φ is arbitrary, we find, for the $\varphi(0)$ terms,

$$-\partial_\alpha(f^\alpha(x)\,\pi(x)) + \frac{1}{2}\partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)) = 0,$$

and for $\partial\varphi(0)$ terms,

$$-f^\alpha(x)\,\pi(x) + \frac{1}{2}\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)) = 0.$$

The $\partial\partial\varphi(0)$ terms vanishes automatically. Altogether, we find the detailed balance condition for Langevin process to be

$$f^\alpha(x)\,\pi(x) = \frac{1}{2}\,\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)). \tag{3.18}$$

Comparing with the stationary solution of Langevin process (equation 3.17), the source-free vector field ν is absent here. Recall in section 2.4 where detailed balance condition was first encountered, we said that detailed balance condition is stronger than just being stationary. Now, in Langevin process, this becomes concrete: *detailed balance condition is stronger than stationary condition in the sense that it lacks the source-free degree of freedom that appears in the stationary condition*. The lost degree of freedom is the cost of ensuring that any initial distribution will finally relax to the stationary.

Chapter 4
Least-Action Principle

We apply the discussions in chapter 3 to least-action principle.

4.1 Conventions in This Chapter

Follow the conventions in chapter 3 (except for section 4.4 where the alphabet $\mathcal{X}$ can be discrete). In addition, we use $P(\theta)$ for a parameterized distribution, where θ is the collection of parameters. Its density function is $p(x, \theta)$, where random variable X takes the value x.

4.2 A Brief Review of Least-Action Principle in Classical Mechanics

In physics, least-action principle gives the dynamics of the state of an evolutionary system, determining how it evolves with time. The state of an evolutionary system is called a **configuration**. As the state changes with time, the evolution of configuration can be seen as a path in a space, like a contrail in the sky, indicating the movement of an airplane. This space is called **configuration space**, which is generally Euclidean, $\mathbb{R}^d$ for some n. A **path** is a function with single parameter $x\colon [t_i, t_f] \to \mathbb{R}^d$, where t_i and t_f denote the initial and final time respectively. Without losing generality, we standardize the time interval from $[t_i, t_f]$ to $[0, 1]$. To introduce the least-action principle, consider the collection of paths with fixed boundaries, that is, $\mathcal{P}(x_0, x_1) := \{x\colon [0, 1] \to \mathbb{R}^d | x(0) = x_0, x(1) = x_1\}$ given the boundaries (x_0, x_1). An **action** is a scalar functional of path with fixed boundaries, thus an action $S(\cdot|x_0, x_1)\colon \mathcal{P}(x_0, x_1) \to \mathbb{R}$, where we use a vertical line to separate variables and those that are given as constants (the boundaries (x_0, x_1)), which should not be confused with the vertical line in conditional probability, like $p(x|y)$. For example, the configuration space of an (one-dimensional) harmonic oscillator is $\mathbb{R}$, and the evolution is characterized by a path $x\colon [0, 1] \to \mathbb{R}$. The action of harmonic oscillator is given by the functional

$$S_{\mathrm{HO}}(x|x_0, x_1) = \frac{1}{2}\int_0^1 \mathrm{d}t\, [\dot{x}^2(t) - \omega^2 x^2(t)], \tag{4.1}$$

where $\dot{x} := \mathrm{d}x/\mathrm{d}t$, $\omega \in \mathbb{R}$, and $x(0) = x_0$, $x(1) = x_1$.

Roughly, least-action principle states that, in the real world, the paths with the fixed boundaries are those that minimize the action. To quantitatively declare the least-action principle, we have to describe the minimum of an action mathematically. Recall that a local minimum, or generally an extremum, $x_\star$ of a function f is characterized by $(\partial f/\partial x^\alpha)(x_\star) = 0$ for each component α. How can we generalize this from function to functional (action is a functional)? The trick is discretizing the time. Precisely, we uniformly separate the time interval $[0, 1]$ into T fragments. Thus, the path x is discretized as a vector $(x(0), x(1/T), \ldots, x((T-1)/T), x(1))$, each component is an endpoint of a fragment. Since the boundaries are fixed in least-action principle, $x(0)$ and $x(1)$ are constant rather than variables. Hence, the true degree of freedom is $(x(1/T), \ldots, x((T-1)/T))$. **Least-action principle in classical mechanics** then states that, given the (discretized) action S and the boundaries (x_0, x_1), there is at most one path $x_\star \in \mathcal{P}(x_0, x_1)$ such that

$$\frac{\partial S}{\partial x(i/T)}(x_\star|x_0, x_1) = 0, \tag{4.2}$$

for each $i=1,\dots,T-1$ and any $T>1$, and that $x_\star$ is the path in real world.

Take harmonic oscillator as example. To discretize its action (equation 4.1), we replace the integral $\int_0^1 \mathrm{d}t$ by mean $(1/T)\sum_{i=0}^{T}$ and $x(t)$ by $x(i/T)$. Thus the second term becomes $(\omega^2/2T)\sum_{i=0}^{T} x^2(i/T)$. For the first term, the derivative $\dot{x}(t)$ is replaced by its difference $T\,[x((i+1)/T)-x(i/T)]$, hence the summation shall terminated at $T-1$ instead of T. Altogether, the action 4.1 is discretized as

$$S_{\mathrm{HO}}(x|x_0,x_1)=\frac{T}{2}\sum_{i=0}^{T-1}\,[x((i+1)/T)-x(i/T)]^2-\frac{\omega^2}{2T}\sum_{i=0}^{T}x^2(i/T),$$

Given i, $x(i/T)$ appears in two terms in S_{HO}, the i and $i+1$ terms in the summation. They have derivatives $T\,[-x((i+1)/T)+x(i/T)]-(\omega^2/T)\,x(i/T)$ and $T[x(i/T)-x((i-1)/T)]$ respectively. So, we find

$$T\frac{\partial S_{\mathrm{HO}}}{\partial x(i/T)}(x_\star|x_0,x_1)=T^2\,[x_\star((i+1)/T)-2x_\star(i/T)+x_\star((i-1)/T)]+\omega^2\,x_\star(i/T),$$

for $i=1,\dots,T-1$. The right hand side is the discretized $\ddot{x}_\star(t)+\omega^2\,x_\star(t)$, for $t\in(0,1)$ (notice we have excluded the $t=0,1$, corresponding to $i=0,T$ respectively). So, least-action principle, $\partial S_{\mathrm{HO}}/\partial x(i/T)(x_\star|x_0,x_1)=0$, implies the correct dynamics of harmonic oscillator in textbooks, which is $\ddot{x}_\star(t)+\omega^2\,x_\star(t)=0$.[4.1]

We can generalize the least-action principle to any system, evolutionary or not, where variables locate in a high-dimensional Euclidean space and, given some conditions, action is a scalar function on it. It states that the real world datum locates in the minimum of the action. Precisely, given the conditioned action S (we may hide the condition y into S instead of explicitly writing it out), there is a at most one $x_\star$ such that

$$\frac{\partial S}{\partial x^\alpha}(x_\star)=0, \tag{4.3}$$

and that $x_\star$ is the real world datum.

There are, however, redundant degrees of freedom in action S. We may construct multiple actions all satisfying equation 4.3. Knowing the extremum of a function cannot imply the shape of the function. The action has much more degrees of freedom than that is needed for revealing the real world datum in classical mechanics. But combined with uncertainty, as we will see in section 4.3, action is completely determined by the real world distribution (the correspondence of real world datum), with nothing redundant.

4.3 Least-Action Principle of Distribution Has No Redundancy

Dynamics in classical mechanics are always deterministic. That is, once the initial conditions (for initial value problem) or the boundaries (for boundary value problem) are fixed, then the path is fully determined. Randomness is forbidden. There are, however, many phenomena in nature that have *intrinsic* randomness. For example, molecular movement obeys a normal distribution with variance proportional to time interval. The dynamics of starling flocks also has intrinsic randomness, which is the "free will" of each bird, so is ant colony, human society, and any interactive system in which each element has some level of intrinsic uncertainty. For these cases, the real world datum is not simply a path, but a distribution of path. In this section, we are to describe these phenomena in the language of Markovian process.

4.1. The dynamics with fixed boundaries is called **boundary value problem**. But in physics, the dynamics we obtained from the least-action principle is applied to **initial value problem**, where the initial "phase" (for physical system, it involves initial position and velocity), instead of boundaries, is fixed. This mysterious application leads to some interesting results. For an mth-order dynamics (for example, harmonic oscillator is a second order dynamics since it involves at most the second derivative of path), an initial value problem has $(T+1-m)$ variables (there are $T+1$ endpoints on the path), since the m degree of freedom has been assigned to the initial values. On the other hand, the boundary value problem has $(T+1-2)$ degree of freedom, since there are always two boundaries ($t=0$ and $t=1$). So, for the success of this mysterious application, we must have $m=2$. That is, the initial value problem has to be second order.

Temporally, we go back to use the old fashioned notation for conditional density functions instead of the arrowed, thus $q_{\Delta t}(x \to y)$ is written as $q_{\Delta t}(y|x)$. By repeatedly applying (discrete time) master equation 2.5, we get

$$p(x_N, N\Delta t) = \int_{\mathbb{R}^d} \mathrm{d}x_0 \cdots \int_{\mathbb{R}^d} \mathrm{d}x_{N-1}\, p(x_0, 0) q_{\Delta t}(x_1|x_0) \cdots q_{\Delta t}(x_N|x_{N-1}). \tag{4.4}$$

The right hand side can be viewed as marginalizing the random variables $(X_0, \ldots, X_{N-1})$, and the product $q_{\Delta t}(x_1|x_0) \cdots q_{\Delta t}(x_N|x_{N-1})$ can be seen as the density function of $(X_1, \ldots, X_N)$, where we have omitted the subscript Δt for simplicity. To see this clearly, we first notice that $q(x_2|x_1) = q(x_2|x_0, x_1)$ holds for any x_0, because $q(x_2|x_1)$ is not explicitly dependent on x_0. Then, we have $q(x_2|x_1)q(x_1|x_0)q(x_0) = q(x_2|x_0, x_1)q(x_1|x_0)q(x_0)$. Repeatedly using the definition of conditional density, it becomes $q(x_2|x_0, x_1)\ q(x_0, x_1) = q(x_0, x_1, x_2)$. Dividing $q(x_0)$ on both sides, we get $q(x_1, x_2|x_0) = q(x_2|x_1)q(x_1|x_0)$. Repeating this step, we will find

$$q(x_N|x_{N-1}) \cdots q(x_1|x_0) = q(x_1, \ldots, x_N|x_0),$$

recognized as a conditional density of random variables $(X_1,..., X_N)$ given x_0. If we regard the series $(x_1,..., x_N)$ as a "movie" or a "path" of evolution of the stochastic system, in which each x_i can be seen as a "frame", then the density function $q(x_1, \ldots, x_N|x_0)$ characterizes the distribution of evolution.

If define $S(x_0, x_1, \ldots, x_N) := -\ln q(x_1, \ldots, x_N|x_0)$, then the master equation becomes

$$p(x_N, N\Delta t) = \int_{\mathbb{R}^d} \mathrm{d}x_0 \cdots \int_{\mathbb{R}^d} \mathrm{d}x_{N-1}\, p(x_0, 0) \exp(-S(x_0, \ldots, x_N)).$$

So, the $[\cdots]$ part transits the density function $p(\cdot, 0)$ to $p(\cdot, N\Delta t)$. It is an integration over all possible ways of evolution. If we treat $(x_0, \ldots, x_N)$ together as a single x, then we can generalize this to any density function $q(x)$ with $x \in \mathbb{R}^d$. We can always define

$$S(x) := -\ln q(x). \tag{4.5}$$

Thus, $q(x) = \exp(-S(x))$.

The S defined by 4.5 has some properties that can be analog to the action in classical mechanics. Indeed, by plugging in the definition of S, we find

$$\int_{\mathbb{R}^d} \mathrm{d}x\, q(x) \frac{\partial S}{\partial x^\alpha}(x) = -\int_{\mathbb{R}^d} \mathrm{d}x\, q(x) \frac{\partial}{\partial x^\alpha} \ln q(x) = -\int_{\mathbb{R}^d} \mathrm{d}x \frac{\partial}{\partial x^\alpha} q(x).$$

The integrand of the right most expression is a divergence, so it results in a boundary integral. But since q, as a density function, is normalized, the boundary integral shall vanish as $\|x\| \to +\infty$. So, we conclude that

$$\mathbb{E}_Q\left[\frac{\partial S}{\partial x^\alpha}\right] = 0.$$

This is analog to equation 4.3, where the minimum $x_\star$ is replaced by the expectation $\mathbb{E}_Q$. Secondly, the distribution Q (whose density function is the q) most likely samples (recall section 1.2 that distribution has a sampler) the x that maximizes q, thus minimizes S. For these reasons, we illustrate the S defined by q as the action of Q. Contrary to the action in classical mechanics, the S here is completely determined by the real world distribution Q (because it is defined by the density function q), without any redundancy. This is the direct implication that distribution involves more information than its most likely datum.

4.4 Data Fitting Is Equivalent to Least-Action Principle of Distribution

Given a collection of real world data, we are to find a distribution that fits the data. These data can be seen as samples from an unknown distribution which characterizes the real world. We are to figure out a method to fit the real world distribution by given some samples of it.

Let $P(\theta)$ represent a distribution parametrized by $\theta \in \mathbb{R}^m$. Its alphabet $\mathcal{X}$ can be a discrete set or a continuous space like $\mathbb{R}^d$. From its density function, $p(\cdot, \theta)$, we get a parameterized action $S(\cdot, \theta)$ such that

$$p(x, \theta) = \exp(-S(x, \theta)) / Z(\theta), \tag{4.6}$$

where $Z(\theta) := \int \mathrm{d}x \exp(-S(x, \theta))$ for ensuring $\int \mathrm{d}x p(x, \theta) = 1$. This is consistent with the action defined by equation 4.5, except that the action here is parameterized, and that we add a normalization factor $Z(\theta)$ for convenience (even though it can be absorbed into $S(x, \theta)$).

What we have is a collection of data, sampled from an unknown distribution Q. And we are to adjust the parameters θ so that $P(\theta)$ approximates Q. To do so, we minimize the relative entropy between Q and $P(\theta)$, which is defined as $H(Q, P(\theta)) := \int \mathrm{d}x q(x) \ln (q(x) / p(x, \theta))$. This expression is formal. Since we do not know the density function of Q, all that we can do with Q is computing the expectation $\mathbb{E}_Q[f] = (1/|Q|) \sum_{x \in Q} f(x)$ for any function f, where we use Q as a set of data. With this realization, we have, after plugging equation 4.6 into $H(Q, P(\theta))$,

$$H(Q, P(\theta)) = \mathbb{E}_Q[\ln q] + \mathbb{E}_Q[S(\cdot, \theta)] + \ln Z(\theta).$$

By omitting the θ-independent terms, we get the loss function

$$L(\theta) := \mathbb{E}_Q[S(\cdot, \theta)] + \ln Z(\theta).$$

The parameters that minimize $L(\theta)$ also minimize $H(Q, P(\theta))$, and vice versa. We can find the $\theta_\star := \operatorname{argmin} L$ by iteratively updating θ along the direction $-\partial L / \partial \theta$. To calculate $-\partial L / \partial \theta$, we start at

$$-\frac{\partial L}{\partial \theta^\alpha}(\theta) = -\mathbb{E}_Q\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right] - \frac{1}{Z(\theta)} \frac{\partial Z}{\partial \theta^\alpha}(\theta).$$

The first term is recognized as $-\mathbb{E}_Q[\partial S / \partial \theta^\alpha]$. For the second term, by inserting the definition of $Z(\theta)$, we get

$$-\frac{1}{Z(\theta)} \frac{\partial Z}{\partial \theta^\alpha}(\theta) = \int_{\mathcal{X}} \mathrm{d}x \frac{\exp(-S(x, \theta))}{Z(\theta)} \frac{\partial S}{\partial \theta^\alpha}(x, \theta) = \int_{\mathcal{X}} \mathrm{d}x \, p(x, \theta) \frac{\partial S}{\partial \theta^\alpha}(x, \theta),$$

where in the last equality, we used the definition of $p(x, \theta)$ (the green factor). This final expression is just the $\mathbb{E}_{P(\theta)}[\partial S / \partial \theta^\alpha]$. Altogether, we arrive at

$$-\frac{\partial L}{\partial \theta^\alpha}(\theta) = \mathbb{E}_{P(\theta)}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right] - \mathbb{E}_Q\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right]. \tag{4.7}$$

At the minimum, we shall have $\partial L / \partial \theta = 0$. Then, we find that $\theta_\star$ obeys

$$\mathbb{E}_{P(\theta_\star)}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta_\star)\right] = \mathbb{E}_Q\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta_\star)\right]. \tag{4.8}$$

It can be read from equation 4.7 that minimizing L is to increase $S(\cdot, \theta)$ on the sampled points (the first term) while decrease it on data points (the second term). As figure 4.1 illustrates, this way of optimization will site real world data onto local minima of $S(\cdot, \theta)$, *in statistical sense*.

The parameterized action S can be constructed out of universal functions such as neural networks. Then, iterating by equation 4.7 until $\|\partial L / \partial \theta\|$ has been small enough gives an $S(\cdot, \theta_\star)$ that mimics the stochastic dynamics of starling flocks. To compute the expectation $\mathbb{E}_{P(\theta)}[\dots]$ in equation 4.7, we employ Monte-Carlo simulation. The transition rate shall satisfy detailed balance condition with $P(\theta)$ as the stationary distribution. For discrete random variables, Monte-Carlo simulation with Metropolis-Hastings algorithm (section 2.7) is available; and for continuous random variables, Langevin process (section 3.6, 3.9, and 5.3) will be more efficient. In addition, many numerical computation modules have build-in samplers for many distributions, such as categorical

distribution and normal distribution; when $P(\theta)$ is happen to be one of them, computing $\mathbb{E}_{P(\theta)}[\dots]$ will be straight-forward.

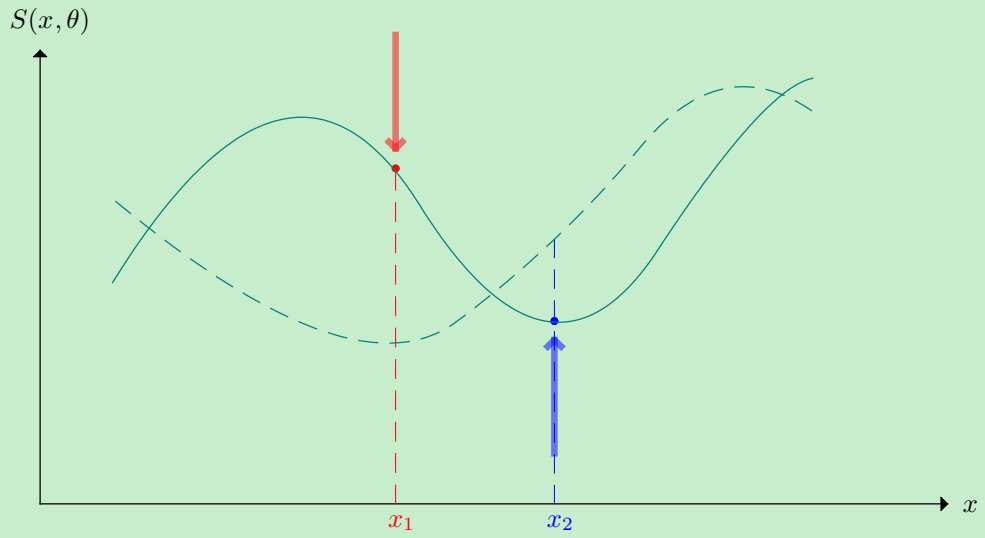

Figure 4.1. This figure illustrate how $\min_\theta L(\theta)$ will site a real world datum onto a local minimum of $S(\cdot,\theta)$. The green curve represents the current not-yet-optimized $S(\cdot,\theta)$. The x_1 (red point) is a real world datum while x_2 (blue point), which is currently a local minimum of $S(\cdot,\theta)$, is not. Minimizing L by tuning θ pushes the $\mathbb{E}_Q[S(\cdot,\theta)]$ down to lower value, corresponding to the red downward double-arrow on x_1. Also, since x_2 is a local minimum, the data points sampled from $p(x,\theta) \propto \exp(-S(x,\theta))$ will accumulate around x_2. So, minimizing L also pulls the $\mathbb{E}_{P(\theta)}[S(\cdot,\theta)]$ up to greater value, corresponding to the blue upward double-arrow on x_2. Altogether, it makes x_1 a local minimum of $S(\cdot,\theta)$, and $S(\cdot,\theta)$ is optimized to be the dashed green curve.

As an example, if we want to get the action that characterizes the stochastic dynamics of starling flocks, we take movies for many flocks. Each movie is a series of frames that log the positions of each bird at each time instant. These movies provide the real world data, namely the distribution Q.

We could construct the parameterized action S by a neural network that outputs a scalar float value. But we can go deeper, assuming that the stochastic dynamics of flocks obeys a Langevin process with constant covariance. Namely, we employ the action 5.10, and construct the f therein using neural network, which outputs a float vector (array).

To evaluate the $\mathbb{E}_{P(\theta)}[\dots]$ in equation 4.7, we first randomly initialize a group of samples, each representing a movie of a flock. Then, we use Langevin process to efficiently simulate master equation. By section 3.7, we iterate each sample by

$$x^\alpha \leftarrow x^\alpha - \nabla^\alpha S(x,\theta)\,\Delta t + \Delta W^\alpha. \tag{4.9}$$

for $0 < \Delta t \ll 1$. The gradient ∇ is taken on x, and the random variable ΔW^α obeys a normal distribution with zero mean and variance Δt. This specific transition density satisfies detailed balance condition (see section 3.9, especially the equation 3.18). Then, theorem 2.1 claims that the iteration 4.9 will relax the samples towards the stationary distribution, which has density function proportional to $\exp(-S(x,\theta))$. There is not need to wait until the samples have been fully relaxed. In practice, we find that ten or several more iterations has been sufficient for a good approximation. Finally, out of these iterated samples, $\mathbb{E}_{P(\theta)}[\cdots]$ is evaluated. Repeating this process and iterating the θ using gradient descent, we arrive at the best-fit $\theta_\star$ which encodes the stochastic dynamics of flocks.

4.5 ♣ Example: Least-Action Principle in Supervised Machine Learning

In this section, we apply the result in section 4.4 to supervised machine learning, including both regression and classification tasks. We suppose that readers who read this section have the basic familiarity to machine learning.

As usual, dataset is characterized by empirical distribution Q. In supervised machine learning, each datum is a pair consisting of an input and an output. By marginalizing the output, we get the input distribution Q_X. Then, dividing Q by Q_X gives the conditional distribution $Q_{Y|X}$, which samples the output by giving an input.

The task of supervised machine learning is searching for an analytical distribution that approximates the $Q_{Y|X}$. To do so, we use a parameterized distribution $P(x,\theta)$, where θ denotes the parameter, and find the best-fit parameter $\theta_\star$ by minimizing the relative entropy $H(P(x,\theta), Q_{Y|X}(x))$, where x is sampled from Q_X. Then, we have the loss function

$$L(\theta) = \mathbb{E}_{x\sim Q_X}[H(P(x,\theta), Q(x))].$$

For regression tasks, the alphabet is $\mathbb{R}$, and $P(x,\theta)$ is a normal distribution with unit variance. Thus, the density function $p(y|x,\theta)$ is

$$p(y|x,\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(h(x,\theta)-y)^2\right)$$

for some scale function $h(x,\theta)$. We read out the action in section 4.4 as

$$S(x,y,\theta) = \frac{1}{2}(h(x,\theta)-y)^2$$

and the normalization factor $Z(x,\theta)$ is constant. So, omitting the θ-independent terms in $L(\theta)$, we get

$$L(\theta) = \mathbb{E}_{x\sim Q_X}[\mathbb{E}_{y\sim Q_{Y|X}(x)}[S(x,y,\theta)]] = \mathbb{E}_{(x,y)\sim Q}[S(x,y,\theta)].$$

For classification tasks, the alphabet is $\{1,\dots,C\}$ for some positive integer C, and $P(x,\theta)$ is a categorical distribution. Thus, the density function $p(y|x,\theta)$ is a softmax function of $h(x,\theta)$. That is,

$$p(\alpha|x,\theta) = \frac{\exp(h^\alpha(x,\theta))}{\sum_{\beta=1}^C \exp(h^\beta(x,\theta))}.$$

We read out the action in section 4.4 as

$$S(x,\alpha,\theta) = -h^\alpha(x,\theta)$$

and the normalization factor $Z(x,\theta) = \sum_\beta \exp(h^\beta(x,\theta))$. The loss function is evaluated to be

$$L(\theta) = -\mathbb{E}_{(x,\alpha)\sim Q}\left[\ln \frac{\exp(h^\alpha(x,\theta))}{\sum_{\beta=1}^C \exp(h^\beta(x,\theta))}\right],$$

which is the usual cross-entropy loss in classification tasks. As discussed in section 4.4, for minimizing the loss function, we have to evaluate the expectation $\mathbb{E}_{P(x,\theta)}[\partial S/\partial\theta]$ by sampling from $P(x,\theta)$. This corresponds to the contrastive learning technique used in training models such as `word2vec` in which the number of classes is extremely large.

Chapter 5
Path Integral

In this chapter, we write down the path integral formulation for generic Markovian process. We find that the path integral is taken on an extended space. We then consider the Langevin process as an instance, for which the extended components may be analytically marginalized.

5.1 Conventions in This Chapter

Follow the conventions in chapter 3. Briefly, the alphabet is an Euclidean space $\mathbb{R}^d$. We use Einstein's convention to simplify notations. And notation $p(x \to y)$ is used for denoting the conditional density function $p(y|x)$.

5.2 Markovian Process with Euclidean Alphabet Can Be Formulated as Path Integral

In section 4.3, we have shown how to define an action using distribution. Also in the same section, we found that the master equation can be written as a integral of all possible paths, as

$$p(x_N, N\,\Delta t) = \int_{\mathbb{R}^d} \mathrm{d}x_0 \cdots \int_{\mathbb{R}^d} \mathrm{d}x_{N-1}\, p(x_0, 0) \exp(-S(x_0, \ldots, x_N)),$$

where the action (of distribution) is given by

$$S(x_0, \ldots, x_N) = -\sum_{i=0}^{N-1} \ln q_{\Delta t}(x_i \to x_{i+1}).$$

If we regard the series $(x_0, \ldots, x_N)$ as a path from x_0 to x_N, then it is an integral over all possible paths.

Path integral formulation was found by Paul Dirac in 1933 who was trying to using Lagrangian in quantum mechanism. It was then developed by physicist Richard Feynman and mathematician Mark Kac in 1947.[5.1] Now, path integral is applied not only to quantum field theory, but also many other areas such as stochastic process. Path integral has the general formalism

$$\int_{\mathbb{R}^d} \mathrm{d}x_0 \cdots \int_{\mathbb{R}^d} \mathrm{d}x_{N-1} \exp(-S(x_0, \ldots, x_N)) f(x_0, \ldots, x_N), \tag{5.1}$$

where a series $(x_0, \ldots, x_N)$ is called a "path", the S is called the "action" of path (in section 4.3, we explained why we should call S an action), and the f is an arbitrary function called an "observable". So, we just found a path integral formulation for master equation.

5.1. *On Distributions of Certain Wiener Functionals* by M. Kac, 1947. DOI: 10.2307/1990512.

To obtain the action, we have to evaluate the logarithm of transition density $\ln q_{\Delta t}(x \to y)$ when Δt is small. This, however, cannot be straight-forward since the leading term of $q_{\Delta t}(x \to y)$ is $\delta(x - y)$ which cannot be converted into exponential. But, we can consider its Fourier transformation (introduced in section 3.5), since $\delta(x - y)$, if regarding as a Dirac's delta function, has Fourier coefficient $\exp(-\mathrm{i}k_\alpha(x^\alpha - y^\alpha))$. This suggest us to consider the Fourier transformation of $q_{\Delta t}$, as

$$\hat{q}_{\Delta t}(x, k) := \int_{\mathbb{R}^d} \mathrm{d}\epsilon \exp(-\mathrm{i}k_\alpha \epsilon^\alpha)\, q_{\Delta t}(x \to x + \epsilon).$$

This forces the alphabet to be Euclidean space $\mathbb{R}^d$, because we cannot perform the same thing on Kronecker's delta when the alphabet is discrete, or when the alphabet is continuous but not Euclidean. Roughly, we have $q_{\Delta t}(x \to x + \epsilon) \approx \delta(\epsilon) + r(x, x + \epsilon)\, \Delta t$. Thus, we may expect that

$$\hat{q}_{\Delta t}(x, k) \approx 1 + \hat{r}(x, k)\, \Delta t \approx \exp(\hat{r}(x, k)\, \Delta t),$$

where we have inserted the Fourier transformation of transition rate, $\hat{r}(x, k) = \int \mathrm{d}\epsilon \exp(-\mathrm{i}k_\alpha \epsilon^\alpha) r(x, x + \epsilon)$. It has inverse Fourier transformation

$$q_{\Delta t}(x \to x + \epsilon) = \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d}\, \hat{q}_{\Delta t}(x, k) \approx \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp(\mathrm{i}k_\alpha \epsilon^\alpha + \hat{r}(x, k)\, \Delta t).$$

Plugging this back into equation 4.4, we get a path integral formulation

$$p(x_N, N\,\Delta t) \approx \int D(x, k)\, p(x_0, 0) \exp\left(\sum_{i=0}^{N-1} \{\mathrm{i}(k_i)_\alpha (x_{i+1}^\alpha - x_i^\alpha) + \hat{r}(x_i, k_i) \Delta t\} \right)$$

with abbreviation

$$\int D(x, k) := \int_{\mathbb{R}^d} \mathrm{d}x_0 \int_{\mathbb{R}^d} \mathrm{d}k_0 \cdots \int_{\mathbb{R}^d} \mathrm{d}x_{N-1} \int_{\mathbb{R}^d} \mathrm{d}k_{N-1}.$$

The residue of this approximation is found to be non-trivial. We tackle this in section 5.2.1. As the result, we find

$$p(x_N, N\,\Delta t) = \int D(x, k)\, p(x_0, 0) \exp(-S(x, k)) + o(N\,\Delta t), \tag{5.2}$$

where the action is

$$S(x, k) = -\sum_{i=0}^{N-1} \{\mathrm{i}(k_i)_\alpha (x_{i+1}^\alpha - x_i^\alpha) + \hat{r}(x_i, k_i) \Delta t\}. \tag{5.3}$$

If we Taylor expand $\hat{r}(x, k)$ by k at the origin, then the coefficient is

$$\lim_{k \to 0} \frac{\partial}{\partial k_{\alpha_1}} \cdots \frac{\partial}{\partial k_{\alpha_n}} \int_{\mathbb{R}^d} \mathrm{d}\epsilon \exp(-\mathrm{i}k_\alpha \epsilon^\alpha) r(x, x + \epsilon) = (-\mathrm{i})^n \int_{\mathbb{R}^d} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_n}) r(x, x + \epsilon),$$

which is recognized as $(-\mathrm{i})^n K_n^{\alpha_1 \cdots \alpha_n}(x)$. We meet the moments of transition rate again (it first appears in section 3.2). Thus, we have

$$\hat{r}(x, k) = \sum_{n=1}^{N_{\mathrm{cut}}} \frac{(-\mathrm{i})^n}{n!} K_n^{\alpha_1 \cdots \alpha_n}(x) k_{\alpha_1} \cdots k_{\alpha_n},$$

where the zeroth order term vanishes (thus the summation begins at $n = 1$) because of the property of transition rate $\int \mathrm{d}\epsilon r(x, x + \epsilon) = 0$. As discussed in section 3.2, the summation terminates at N_{cut}. Then, the action becomes

$$S(x, k) = -\sum_{i=0}^{N-1} \Delta t \left\{ \mathrm{i} \left(\frac{x_{i+1}^\alpha - x_i^\alpha}{\Delta t} - K_1^\alpha(x_i) \right) (k_i)_\alpha + \sum_{n=2}^{N_{\mathrm{cut}}} \frac{(-\mathrm{i})^n}{n!} K_n^{\alpha_1 \cdots \alpha_n}(x_i) (k_i)_{\alpha_1} \cdots (k_i)_{\alpha_n} \right\}. \tag{5.4}$$

The path is a poly-line on $\mathbb{R}^{2d}$ (involving both $x\in\mathbb{R}^d$ and $k\in\mathbb{R}^d$).[5.2] But in a specific situation where $N_{\text{cut}}=2$ and $K_2(x)$ is constant (matrix), the k-components can be analytically marginalized (section 5.3). Then, the path integral is taken on the alphabet $\mathbb{R}^d$ (namely, the path is a series $(x_0,x_1,\dots,x_N)$ with each $x_i\in\mathbb{R}^d$).

5.2.1 ♣ Estimation of the Residue

To explicitly evaluate the residue left in section 5.2, we have to consider the full expansion of $q_{\Delta t}$ (equation 2.6), rather than expanding to the first order of Δt. We have evaluated the zeroth and the first order coefficients in the full expansion. Now for higher orders, define

$$\hat{g}_n(x,k):=\int_{\mathbb{R}^d}\mathrm{d}\epsilon\exp(-\mathrm{i}k_\alpha\epsilon^\alpha)\int_{\mathbb{R}^d}\mathrm{d}y_1\cdots\int_{\mathbb{R}^d}\mathrm{d}y_{n-1}\,r(x,y_1)r(y_1,y_2)\cdots r(y_{n-1},x+\epsilon),$$

for $n\geqslant 1$. Then, the full expansion of $q_{\Delta t}$ gives

$$\hat{q}_{\Delta t}(x,k)=1+\sum_{n=1}^{+\infty}\frac{\Delta t^n}{n!}\hat{g}_n(x,k).$$

The residue hides in the $\hat{g}_n(x,k)$s. We are to calculate $\hat{g}_n(x,k)$ by induction, starting at $\hat{g}_1(x,k)=\hat{r}(x,k)$. Directly,

$$\hat{g}_{n+1}(x,k)=\int_{\mathbb{R}^d}\mathrm{d}y_1 r(x,y_1)\left[\int_{\mathbb{R}^d}\mathrm{d}y_2\cdots\int_{\mathbb{R}^d}\mathrm{d}y_n\int_{\mathbb{R}^d}\mathrm{d}\epsilon\exp(-\mathrm{i}k_\alpha\epsilon^\alpha)r(y_1,y_2)\cdots r(y_n,x+\epsilon)\right].$$

Denoting $\epsilon':=\epsilon+x-y_1$, we find

$$\begin{aligned}\hat{g}_{n+1}(x,k)=&\int_{\mathbb{R}^d}\mathrm{d}y_1\exp(-\mathrm{i}k_\alpha(y_1^\alpha-x^\alpha))r(x,y_1)\times\\&\times\left[\int_{\mathbb{R}^d}\mathrm{d}y_2\cdots\int_{\mathbb{R}^d}\mathrm{d}y_n\int_{\mathbb{R}^d}\mathrm{d}\epsilon'\exp(-\mathrm{i}k_\alpha\epsilon'^\alpha)r(y_1,y_2)\cdots r(y_n,x+\epsilon)\right]\end{aligned}$$

The $[\cdots]$ factor is recognized as $\hat{g}_n(x,k)$. Thus, (reuse the ϵ notation for simplicity)

$$\begin{aligned}\hat{g}_{n+1}(x,k)&=\int_{\mathbb{R}^d}\mathrm{d}y_1\hat{g}(y_1,k)\exp(-\mathrm{i}k_\alpha(y_1^\alpha-x^\alpha))r(x,y_1)\\&=\int_{\mathbb{R}^d}\mathrm{d}\epsilon\,\hat{g}(x+\epsilon,k)\exp(-\mathrm{i}k_\alpha\epsilon^\alpha)r(x,x+\epsilon).\end{aligned}$$

Then, Taylor expanding $\hat{g}(y_1+\epsilon,k)$ by ϵ at origin results in (∂ is derived on the first variable, and ∂' on the second)

$$\hat{g}_{n+1}(x,k)=\sum_{m=0}^{+\infty}\frac{1}{m!}\frac{\partial^m\hat{g}_n}{\partial x^{\alpha_1}\cdots\partial x^{\alpha_m}}(x,k)\int_{\mathbb{R}^d}\mathrm{d}\epsilon\,(\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_m})\exp(-\mathrm{i}k_\alpha\epsilon^\alpha)\,r(x,x+\epsilon).$$

The integral is recognized as partial derivatives $\partial^m\hat{r}\,/\,[\partial(-\mathrm{i}k_{\alpha_1})\cdots\partial(-\mathrm{i}k_{\alpha_m})]$, thus we arrive at

$$\hat{g}_{n+1}(x,k)=\sum_{m=0}^{+\infty}\frac{\mathrm{i}^m}{m!}\frac{\partial^m\hat{g}_n}{\partial x^{\alpha_1}\cdots\partial x^{\alpha_m}}(x,k)\frac{\partial^m\hat{r}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_m}}(x,k).\tag{5.5}$$

In $\hat{g}_n(x,k)$, the term with $m=0$ is $[\hat{r}(x,k)]^n$. So, we conclude that

$$\hat{q}_{\Delta t}(x,k)=1+\sum_{n=1}^{+\infty}\frac{\Delta t^n}{n!}\hat{g}_n(x,k)=\exp(\hat{r}(x,k)\,\Delta t)+\hat{\zeta}(x,k,\Delta t),$$

where $\hat{\zeta}(x,k,\Delta t)$ collects the terms in each $\hat{g}_n(x,k)$ except for the $[\hat{r}(x,k)]^n$. Together with equation 5.5, we get

$$\hat{\zeta}(x,k,\Delta t)=\sum_{n=2}^{+\infty}\frac{\Delta t^n}{n!}\sum_{m=1}^{+\infty}\frac{\mathrm{i}^m}{m!}\frac{\partial^m\hat{g}_{n-1}}{\partial x^{\alpha_1}\cdots\partial x^{\alpha_m}}(x,k)\frac{\partial^m\hat{r}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_m}}(x,k).\tag{5.6}$$

5.2. Analogy to classical mechanics, the k-components can be seen as momenta. Then, the $\mathbb{R}^{2d}$ is regarded as a "phase space".

Since the terms like $[\hat{r}(x,k)]^n$ have been absent in $\hat{\zeta}(x,k,\Delta t)$, m starts at 1. And since $\hat{g}_1(x,k)=\hat{r}(x,k)$, it contributes nothing to $\hat{\zeta}(x,k,\Delta t)$, and n starts at 2. Thus, $\hat{\zeta}(x,k,\Delta t)=o(\Delta t)$ and it is the residue we are seeking for. Temporally, we have no idea whether the series in $\hat{\zeta}(x,k,\Delta t)$ converge or not, but keep them formal (we will examine this later).

For going back to $q_{\Delta t}(x\to x+\epsilon)$, we perform inverse Fourier transformation. This refers to the expansion of a function f as[5.3]

$$f(x)=\sum_{n=0}^{+\infty}\frac{(-\mathrm{i})^n}{n!}(\partial^{\alpha_1}\cdots\partial^{\alpha_n}\hat{f})(0)(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x). \tag{5.7}$$

Notice that the left hand side of this expansion is a function, while the right hand side is a series of generalized functions (the derivatives of Dirac's delta function). So, this expansion has meaning only when it is applied onto a test function in Schwartz space. This is just our situation, where $q_{\Delta t}$ is applied to its right hand side in equation 4.4, which, as we have discussed in the end of section 3.3, is in Schwartz space if initially $p(\cdot,0)$ is. Replacing $\hat{f}(k)$ by $\hat{\zeta}(x,k,\Delta t)$, we find

$$\zeta(x,\epsilon,\Delta t)=\sum_{l=0}^{+\infty}\frac{(-\mathrm{i})^l}{l!}\frac{\partial^l\hat{\zeta}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_l}}(x,0,\Delta t)(\partial_{\alpha_1}\cdots\partial_{\alpha_l}\delta)(\epsilon).$$

Applying to $\varphi\in S(\mathbb{R}^d)$ and taking integration by parts gives

$$\int_{\mathbb{R}^d}\mathrm{d}\epsilon\,\zeta(x,\epsilon,\Delta t)\varphi(\epsilon)=\sum_{l=0}^{+\infty}\frac{\mathrm{i}^l}{l!}\frac{\partial^l\hat{\zeta}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_l}}(x,0,\Delta t)(\partial_{\alpha_1}\cdots\partial_{\alpha_l}\varphi)(0).$$

So, we have to show that the coefficients are well-defined. Inserting equation 5.6, we find the coefficient

$$\frac{\partial^l\hat{\zeta}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_l}}(x,0,\Delta t)=\sum_{n=2}^{+\infty}\frac{\Delta t^n}{n!}\sum_{m=1}^{+\infty}\frac{\mathrm{i}^m}{m!}\frac{\partial}{\partial k_{\alpha_1}}\cdots\frac{\partial}{\partial k_{\alpha_l}}\left[\frac{\partial^m\hat{g}_{n-1}}{\partial x^{\alpha_1}\cdots\partial x^{\alpha_m}}(x,0)\frac{\partial^m\hat{r}}{\partial k_{\alpha_1}\cdots\partial k_{\alpha_m}}(x,0)\right].$$

5.3. This can be viewed as a generalization of Kramers-Moyal expansion. In fact, we are to prove this expansion by following the steps in section 3.3, in which we proved Kramers-Moyal expansion. Explicitly, consider a function $\varphi\in S(\mathbb{R}^d)$, thus Taylor expanding φ at origin gives

$$\int_{\mathbb{R}^d}\mathrm{d}x\,f(x)\,\varphi(x)=\sum_{n=0}^{+\infty}\frac{1}{n!}\left[\int_{\mathbb{R}^d}\mathrm{d}x\,f(x)\,(x^{\alpha_1}\cdots x^{\alpha_n})\right](\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0).$$

We relate the integral in the $[\cdots]$ to the Fourier transformation $\hat{f}$ by

$$(\partial^{\alpha_1}\cdots\partial^{\alpha_n}\hat{f})(0)=\lim_{k\to 0}\frac{\partial}{\partial k_{\alpha_1}}\cdots\frac{\partial}{\partial k_{\alpha_n}}\int_{\mathbb{R}^d}\mathrm{d}x\exp(-\mathrm{i}k_\alpha\epsilon^\alpha)\,f(x)=(-\mathrm{i})^n\int_{\mathbb{R}^d}\mathrm{d}x\,f(x)\,(x^{\alpha_1}\cdots x^{\alpha_n}).$$

Thus,

$$\int_{\mathbb{R}^d}\mathrm{d}x\,f(x)\,\varphi(x)=\sum_{n=0}^{+\infty}\frac{\mathrm{i}^n}{n!}(\partial^{\alpha_1}\cdots\partial^{\alpha_n}\hat{f})(0)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0).$$

On the other hand, because of the identity

$$(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0)=\int_{\mathbb{R}^d}\mathrm{d}x\,\delta(x)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(x),$$

integration by parts on the right hand side gives

$$(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\varphi)(0)=(-1)^n\int_{\mathbb{R}^d}\mathrm{d}x\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x)\,\varphi(x).$$

Plugging this back, we find

$$\int_{\mathbb{R}^d}\mathrm{d}x\,f(x)\,\varphi(x)=\sum_{n=0}^{+\infty}\frac{(-\mathrm{i})^n}{n!}(\partial^{\alpha_1}\cdots\partial^{\alpha_n}\hat{f})(0)\int_{\mathbb{R}^d}\mathrm{d}x\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x)\,\varphi(x).$$

Since φ is arbitrary, we finall arrive at

$$f(x)=\sum_{n=0}^{+\infty}\frac{(-\mathrm{i})^n}{n!}(\partial^{\alpha_1}\cdots\partial^{\alpha_n}\hat{f})(0)\,(\partial_{\alpha_1}\cdots\partial_{\alpha_n}\delta)(x).$$

The terms in the series of the right hand side are all proportional to partial derivatives of $\hat{r}(x,k)$ on k at origin, which is proportional to the moments of transition rate, $K(x)$. As it is concluded in section 3.2, $K_n(x)$ vanishes for all $n > N_{\text{cut}}$. So, the series also terminates at finite m. In the series of n, the n-th term (as a series of m) approximately terminates at $m = nN_{\text{cut}}$. And since all $K(x)$ are bounded, the series of m increases linearly with n. Hence, the series of n converges. It means that, the application onto φ results in a well-defined residue, proportional to Δt^2.

5.3 Langevin Process with Constant Covariance Has a Path Integral on Alphabet

The path integral 5.2 integrates on paths over both x and k. In this section, we are to see when we can marginalize (integrate out) the k-component, leaving only the x, namely the path on the alphabet.

Given $i \in \{0, \dots, N-1\}$, we are to integrate out the k_i in equation 5.2, together with the action 5.4. Now we can safely neglect the residue, and write the integral as (replacing x_i by x, k_i by k, and $x_{i+1} - x_i$ by ϵ for simplicity)

$$I(x) := \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp\left(\mathrm{i}\left[\epsilon^\alpha - K_1^\alpha(x)\,\Delta t\right]k_\alpha - \frac{\Delta t}{2!} K_2^{\alpha\beta}(x)\, k_\alpha k_\beta + \frac{\mathrm{i}\Delta t}{3!} K_3^{\alpha\beta\gamma}(x)\, k_\alpha k_\beta k_\gamma + \cdots \right). \tag{5.8}$$

The series terminates at the cut-off N_{cut} of K_n.

This integral is complicated except for $N_{\text{cut}} = 2$ where it becomes a Gaussian integral, and the Markovian process deduces to a Langevin process, defined in section 3.6. In this situation, we have (re-denote K_1 by f and K_2 by Σ as in section 3.6)

$$I(x) = \int_{\mathbb{R}^d} \frac{\mathrm{d}k}{(2\pi)^d} \exp\left(\mathrm{i}\left[\epsilon^\alpha - f^\alpha(x)\,\Delta t\right] k_\alpha - \frac{\Delta t}{2} \Sigma^{\alpha\beta}(x)\, k_\alpha k_\beta \right).$$

As proved in the same section, Σ is everywhere positive definite. Then by the formula of Gaussian integral, which holds for any positive definite matrix A,

$$\int_{\mathbb{R}^d} \mathrm{d}x \exp\left(-\frac{1}{2} A_{\alpha\beta}\, x^\alpha\, x^\beta + b_\alpha\, x^\alpha \right) = \sqrt{\frac{(2\pi)^d}{\det A}} \exp\left(\frac{1}{2} (A^{-1})^{\alpha\beta} b_\alpha\, b_\beta \right),$$

we find (replacing $A_{\alpha\beta}$ by $\Sigma^{\alpha\beta}(x)\,\Delta t$ and b_α by $\mathrm{i}\left[\epsilon^\alpha - f^\alpha(x)\,\Delta t\right]$)

$$I(x) = \frac{1}{\sqrt{(2\pi\Delta t)^d \det \Sigma(x)}} \exp\left(-\frac{\Delta t}{2} [\Sigma^{-1}(x)]_{\alpha\beta} \left[\frac{\epsilon^\alpha}{\Delta t} - f^\alpha(x) \right] \left[\frac{\epsilon^\beta}{\Delta t} - f^\beta(x) \right] \right). \tag{5.9}$$

Notice that this is consistent with the result in section 3.7.

But there is an extra factor $\sqrt{\det \Sigma(x)}$ out of exponential. To match the path integral formulism 5.1, in which all integration variables are in the exponential, we have to convert the factor into exponential too. It is found that only when Σ is constant can we do so (we left the general situation to section 5.4). As a real symmetric matrix, we perform the Cholesky factorization introduced in section 3.7 to Σ, so that $\Sigma = C^T C$ for some $d \times d$ matrix C. Coordinate transformation $x \to xC$ (also taken on ϵ and f) then eliminates the Σ matrix in the exponential, leaving

$$I(x) = \exp\left(-\frac{\Delta t}{2} \sum_{\alpha=1}^{d} \left[\frac{\epsilon^\alpha}{\Delta t} - f^\alpha(x) \right]^2 \right),$$

up to a constant term in the exponential. Plugging $I(x)$ back to equation 5.2, we find

$$p(x_N, N\Delta t) = \int \mathrm{D}(x)\, p(x_0, 0) \exp(-S(x) + NC) + o(N\Delta t)$$

with

$$\int D(x) := \int \mathrm{d}x_0 \cdots \int \mathrm{d}x_{N-1}$$

and

$$S(x)=\sum_{i=0}^{N-1}\Delta t\left\{\frac{1}{2}\sum_{\alpha=1}^{d}\left[\left(\frac{x_{i+1}^{\alpha}-x_i^{\alpha}}{\Delta t}-f^{\alpha}(x_i)\right)\right]^2\right\}. \tag{5.10}$$

It is a path integral on the alphabet.

Final remark on cut-off $N_{\text{cut}}=2$. If choose $N_{\text{cut}}>2$, it is hard to see how to integrate the improper integral 5.8, and even to show why it is finite. For example, if $N_{\text{cut}}=4$, the $(\Delta t/4!)K_4^{\alpha\beta\gamma\sigma}(x)\,k_\alpha k_\beta k_\gamma k_\sigma$ term will dominate the integral when k is far from origin. But we cannot ensure that this term will suppress the integrand as $\|k\|\to+\infty$ so as to make the improper integral finite. We cannot even diagonalize the fourth order symmetric tensor $K_4(x)$ (because diagonalizing a fourth order symmetric tensor has $\mathcal{O}(d^4)$ restrictions, but a coordinate transformation has only $\mathcal{O}(d^2)$ degrees of freedom, so this cannot be done except for specific situations).

5.4 ♣ Grassmann Variable, Berezin Integral, and Ghosts

We have to briefly introduce Grassmann variable, on which Berezin integral is based. Grassmann variable is an extension of real (or complex) variable,by introducing in the anti-commutative variables. Given a set of variables $\{\zeta_i|i=1,\dots,n\}$, we demand that the anti-commutative relation between ζs

$$\zeta_i\,\zeta_j=-\zeta_j\,\zeta_i.$$

But for any real (or complex) variable x, we demand a commutative relation

$$x\,\zeta_i=\zeta_i\,x.$$

Because of anti-commutation, we have $\zeta_i\zeta_i=-\zeta_i\zeta_i$, thus $\zeta_i\zeta_i=0$. So, a polynomial of single Grassmann variable is always linear, as $f(\zeta)=a+b\zeta$ where coefficients $a,b\in\mathbb{R}$. And a polynomial of two Grassmann variables, ζ and η, is

$$f(\zeta,\eta)=a+b\zeta+c\eta+d\zeta\eta,$$

where coefficients are real numbers. Extra term like $\zeta\eta\zeta=-\zeta\zeta\eta=\zeta\zeta\eta$, thus vanishes. Generally, a polynomial of n Grassmann variables has terms with no more than n factors.

A function can be defined via its Taylor expansion in real number. So, the exponential function for Grassmann number is defined by

$$\exp(\zeta)=1+\zeta+\frac{1}{2!}\zeta^2+\frac{1}{3!}\zeta^3+\cdots.$$

If ζ is a single Grassmann variable, we have $\exp(\zeta)=1+\zeta$ since other terms are all vanishing. This linearity, however, breaks when consider more Grassmann variables. For example, consider $\exp(\sum_{i=1}^{n}\zeta_i\,\eta_i)$ where ζ and η are multiple Grassmann variables. The maximal order is

$$\frac{1}{n!}\left(\sum_{i=1}^{n}\zeta_i\,\eta_i\right)^n=\zeta_1\eta_1\cdots\zeta_n\eta_n,$$

since $\zeta_i\eta_i$ and $\zeta_j\eta_j$ is commutative.

Now, we introduce the integral on Grassmann variables. The integral is a linear operator, defined by

$$\int \mathrm{d}\zeta_i\,\zeta_j=\delta_{ij}.$$

The integral measures $\{\mathrm{d}\zeta_i|i=1,\dots,n\}$ are also anti-commutative, namely

$$\mathrm{d}\zeta_i\mathrm{d}\zeta_j=-\mathrm{d}\zeta_j\mathrm{d}\zeta_i.$$

So, as an example, we have

$$\int \mathrm{d}\eta_n\int \mathrm{d}\zeta_n\cdots\int \mathrm{d}\eta_1\int \mathrm{d}\zeta_1\exp\left(\sum_{i=1}^{n}\zeta_i\,\eta_i\right)=\cdots+\left[\int \mathrm{d}\eta_n\int \mathrm{d}\zeta_n\cdots\int \mathrm{d}\eta_1\int \mathrm{d}\zeta_1(\zeta_1\eta_1\cdots\zeta_n\eta_n)\right]=1, \tag{5.11}$$

where the $[\cdots]$ terms are all vanishing because they do not have sufficiently many Grassmann variables.

Next, we investigate how linear transformation effects the integral measure. To do so, consider a d-dimensional Grassmann variable, in which each component is a single Grassmann variable, $\zeta = (\zeta^1, \ldots, \zeta^d)$. If we take the linear transformation

$$\zeta'_\alpha = A_{\alpha\beta}\zeta^\beta,$$

then the product

$$\zeta'_1 \cdots \zeta'_d = (A_{1\beta}\zeta^\beta)\cdots(A_{d\beta}\zeta^\beta) = \sum_{\mathrm{perm}(\beta)} A_{1\beta_1}A_{2\beta_2}\cdots A_{d\beta_d}\zeta^{\beta_1}\cdots\zeta^{\beta_d},$$

where the summation includes all permutations of $(\beta_1, \ldots, \beta_d)$. If we re-arrange the $\zeta^{\beta_1}\cdots\zeta^{\beta_d}$ factor to $\zeta^1\cdots\zeta^d$, then there comes a factor $(-1)^{\mathrm{sign}(\beta)}$, where $\mathrm{sign}(\beta)$ is the signature of the permutation $(\beta_1, \ldots, \beta_d)$. For example, $(2, 1, 3, 4, \ldots, d)$ has signature 1 since by only one permutation $1 \leftrightarrow 2$, can we recover the natural order $(1, 2, 3, 4, \ldots, d)$. So,

$$\zeta'_1 \cdots \zeta'_d = \zeta^1 \cdots \zeta^d \times \sum_{\mathrm{perm}(\beta)} (-1)^{\mathrm{sign}(\beta)} A_{1\beta_1}A_{2\beta_2}\cdots A_{d\beta_d}.$$

The right hand side is recognized as $\zeta^1 \cdots \zeta^d \times \det A$. On the other hand, since both $\int \mathrm{d}\zeta_d \cdots \int \mathrm{d}\zeta_1 \, (\zeta_1 \cdots \zeta_d)$ and $\int \mathrm{d}\zeta'_d \cdots \int \mathrm{d}\zeta'_1 \, (\zeta'_1 \cdots \zeta'_d)$ results in one, namely

$$\int \mathrm{d}\zeta'_d \cdots \int \mathrm{d}\zeta'_1 \, (\zeta'_1 \cdots \zeta'_d) = \int \mathrm{d}\zeta_d \cdots \int \mathrm{d}\zeta_1 \, (\zeta_1 \cdots \zeta_d),$$

we find the transformation of integral measure, as

$$\int \mathrm{d}\zeta'_d \cdots \int \mathrm{d}\zeta'_1 = \det A \times \int \mathrm{d}\zeta_d \cdots \int \mathrm{d}\zeta_1. \tag{5.12}$$

After introducing Grassmann variable and the integration on multiple such variables, we come to the formula of Berezin integral, which is an analogy of Gaussian integral for Grassmann variables. Consider the Gaussian-like integral

$$\int \mathrm{d}\zeta \mathrm{d}\eta \exp(A_{\alpha\beta}\zeta^\alpha \eta^\beta),$$

where $A \in \mathbb{R}^{d\times d}$ (or $\mathbb{C}^{d\times d}$) and we have briefly denoted

$$\int \mathrm{d}\zeta \mathrm{d}\eta := \int \mathrm{d}\zeta_d \int \mathrm{d}\eta_d \cdots \int \mathrm{d}\zeta_1 \int \mathrm{d}\eta_1. \tag{5.13}$$

Defining $\zeta'_\beta := A_{\alpha\beta}\zeta^\alpha$ and using equation 5.12, it becomes

$$\int \mathrm{d}\zeta \mathrm{d}\eta \exp(A_{\alpha\beta}\zeta^\alpha \eta^\beta) = \det A \times \int \mathrm{d}\zeta' \mathrm{d}\eta \exp(\zeta'_\beta \eta^\beta).$$

The rightmost factor has been evaluated in equation 5.11, which results in 1, thus

$$\int \mathrm{d}\zeta \mathrm{d}\eta \exp(A_{\alpha\beta}\zeta^\alpha \eta^\beta) = \det A. \tag{5.14}$$

This is called **Berezin integral**, named after the Soviet Russian mathematician and physicist Felix Berezin.

After introducing Berezin integral, we go back to deal with equation 5.9 where Σ is not a constant. We first use Cholesky factorization to remove the square root and the fraction. Then, using Berezin integral to convert the determinant into exponential.

Introduced in section 3.7, Cholesky factorization decomposes $\Sigma(x)$ into $C^T(x)\, C(x)$. Instead of the matrix-valued field C, it is more convenient to use its inverse $R(x) := C^{-1}(x)$, thus $\Sigma^{-1}(x) = R^T(x)\, R(x)$. So, we have

$$[\det \Sigma(x)]^{-1/2} = \det R(x),$$

and thus

$$I(x) = \frac{1}{\sqrt{(2\pi\Delta t)^d}} [\det R(x)] \exp\left(-\frac{\Delta t}{2} \sum_{\alpha=1}^{d} \left[R_{\alpha\beta}(x) \left(\frac{\epsilon^\beta}{\Delta t} - f^\beta(x) \right) \right]^2 \right). \tag{5.15}$$

Now, the determinant gets rid of square root and fraction. Remark that $R(x)$ may not be a symmetric matrix.

Then, using Berezin integral, we can convert the $\det R(x)$ factor in equation 5.15 into exponential. Replacing A by $R(x)\Delta t$, we find

$$\int \mathrm{d}\zeta \mathrm{d}\eta \exp(-\Delta t R_{\alpha\beta}(x)\, \zeta^\alpha \eta^\beta) = \det[R(x)\Delta t] = \Delta t^d \det R(x),$$

we convert the determinant into exponential, as[5.4]

$$q_{\Delta t}(x \to x+\epsilon) = \frac{1}{\sqrt{(2\pi\Delta t^3)^d}} \int \mathrm{d}\zeta \mathrm{d}\eta \exp\left(-\frac{\Delta t}{2} \sum_{\alpha=1}^{d} \left[R_{\alpha\beta}(x) \left(\frac{\epsilon^\beta}{\Delta t} - f^\beta(x) \right) \right]^2 - \Delta t R_{\alpha\beta}(x) \zeta^\alpha \eta^\beta \right).$$

In physics, the Grassmann variables ζ and η are called "ghost variables".

Plugging back to master equation 5.2, we get

$$p(x_N, N\Delta t) = \int \mathrm{D}(x, \zeta, \eta)\, p(x_0, 0) \exp(-S(x, \zeta, \eta) + C) + o(N\Delta t).$$

with the abbreviation

$$\int \mathrm{D}(x, \zeta, \eta) := \int_{\mathbb{R}^n} \mathrm{d}x_0 \cdots \int_{\mathbb{R}^n} \mathrm{d}x_{N-1} \int \mathrm{d}\zeta_0 \mathrm{d}\eta_0 \cdots \int \mathrm{d}\zeta_{N-1} \mathrm{d}\eta_{N-1}$$

and the "action" of Langevin process is

$$S(x, \zeta, \eta) := \sum_{i=0}^{N-1} \Delta t \left\{ \frac{1}{2} \sum_{\alpha=1}^{d} \left[R_{\alpha\beta}(x_i) \left(\frac{x_{i+1}^\beta - x_i^\beta}{\Delta t} - f^\beta(x_i) \right) \right]^2 + R_{\alpha\beta}(x_i) \zeta_i^\alpha \eta_i^\beta \right\}. \tag{5.16}$$

The $C := -(d/2)(\ln 2\pi + 3\ln \Delta t)$ is independent of x, ζ, or η, thus is regarded as constant.

5.5 ♡ Fisher Matrix Characterizes Information Propagation in a Stochastic System

Now in a try autumn day, you stand on the open ground, looking at a starling flock flying under the blue sky. Suddenly, an eagle dives into the flock. Some bird in the flock first notices the danger, trying to avoid by turning direction. Other birds in the neighbor notice the behavior, may follow it too, even though they have not seen the diving eagle yet. Then from neighbors to neighbors, the danger signal may soon spread in the flock. The fact will not be so because each bird has some degree of randomness (or free will). Because the random movement itself is another signal (or noise) to propagate in the flock, it pollutes the danger signal. This randomness, however, is essential for a flock to survive. It assigns flexibility to the flock so that eagles cannot predict the direction it moves. In reality, the connection between the neighbors and the randomness of each individual are properly balanced, so that the flock has sufficient flexibility and a danger signal can propagate far enough within the flock.

This phenomenon appears everywhere in Nature: a group of individuals (such as a starling flock or an ant colony) behaves like an "intelligent" agent, because of the elaborate balance between determinacy and randomness. It is a typical stochastic system that can be described using the techniques we have developed so far.

5.4. You may convert the determinant directly into exponential by using logarithm, namely $\det R(x) = \exp\{\ln[\det R(x)]\}$. This fails in our situation because we expect the term in the exponential to be proportional to Δt.

To characterize the information propagation in a stochastic system, we consider the transition density $q_t(x \to y)$. It describes the probability (or portion) that the system transits from x to y after time t. An initial perturbation $x \to x+\varepsilon$ will affect the distribution after time t. Following section 4.3, the action is defined to be $S_t(x,y) := -\ln q_t(x \to y)$, which also depends on time t and initial state x. Then, the difference in the distribution after time t is given by the relative entropy

$$H(Q_t(x), Q_t(x+\varepsilon)) = \int \mathrm{d}y q_t(x \to y) \ln \frac{q_t(x \to y)}{q_t(x+\varepsilon \to y)} = \mathbb{E}_{y\sim Q_t(x)}[S_t(x+\varepsilon, y) - S_t(x,y)].$$

Taylor expansion by ε at origin gives

$$H(Q_t(x), Q_t(x+\varepsilon)) = \mathbb{E}_{y\sim Q_t(x)}\left[\frac{\partial S_t}{\partial x^\alpha}(x,y)\right]\varepsilon^\alpha + \frac{1}{2}\mathbb{E}_{y\sim Q_t(x)}\left[\frac{\partial^2 S_t}{\partial x^\alpha \partial x^\beta}(x,y)\right]\varepsilon^\alpha\varepsilon^\beta + o(\varepsilon^2).$$

The first term in the right hand side vanishes, since

$$\mathbb{E}_{y\sim Q_t(x)}\left[\frac{\partial S_t}{\partial x^\alpha}(x,y)\right] = -\int_{\mathbb{R}^d} \mathrm{d}y q_t(x \to y) \frac{\partial}{\partial x^\alpha} \ln q_t(x \to y) = -\frac{\partial}{\partial x^\alpha}\int_{\mathbb{R}^d} \mathrm{d}y q_t(x \to y),$$

which vanishes because of normalization $\int \mathrm{d}y q_t(x \to y) = 1$ for any x. For the second term, define

$$\mathcal{F}_{\alpha\beta}(x,t) := \mathbb{E}_{y\sim Q_t(x)}\left[\frac{\partial^2 S_t}{\partial x^\alpha \partial x^\beta}(x,y)\right],$$

then we find

$$H(Q_t(x), Q_t(x+\varepsilon)) = \frac{1}{2}\mathcal{F}_{\alpha\beta}(x,t)\varepsilon^\alpha\varepsilon^\beta + o(\varepsilon^2).$$

The matrix-valued field $\mathcal{F}$ is called **Fisher matrix**, named after the British polymath Ronald Fisher.[5.5] It characterizes the information propagation in a stochastic system.

As an example, consider the Wiener process introduced in section 3.5. Its transition density is

$$q_t(x \to y) = \prod_{\alpha=1}^d \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2t}(x^\alpha - y^\alpha)^2\right).$$

So, it has action $S_t(x,y) = (1/2t)\sum_\alpha (x^\alpha - y^\alpha)^2$, thus its Fisher matrix is evaluated to be $\mathcal{F}_{\alpha\beta}(x, t) = \delta^{\alpha\beta}/t$. It decays with t increasing, indicating that the initial information diminishes during the propagation or evolution.

But for an arbitrary Markovian process, the action $S_t(x,y)$ is difficult to evaluate when t is not small enough.

5.5. If we regard the space where the distribution $Q_t(x)$ lives as a Riemannian surface (that changes with time), then the Fisher matrix $\mathcal{F}$ serves as a metric of the Riemannian surface.

You may argue that relative entropy is not a distance, because it is not symmetric. Indeed, generally $H(Q_t(x), Q_t(x+\varepsilon)) \neq H(Q_t(x+\varepsilon), Q_t(x))$. But, we can consider its symmetric form $D_{\mathrm{JS}}(Q_t(x), Q_t(x+\varepsilon)) := [H(Q_t(x), Q_t(x+\varepsilon)) + H(Q_t(x+\varepsilon), Q_t(x))]/2$, then D_{JS}, named by **Jensen-Shannon distance**, is indeed a distance. It can be shown that $D_{\mathrm{JS}}(Q_t(x), Q_t(x+\varepsilon)) = (1/2)\mathcal{F}_{\alpha\beta}(x,t)\varepsilon^\alpha\varepsilon^\beta + o(\varepsilon^2)$ too.

Epilogue

In the summer of 2024, I was wondering how a stochastic system relaxes to its equilibrium. I did not find a textbook that matched my expectation. Traditional textbooks are either too complicated for me or not rigorous. So, I decided to build up the theory by hands. I found it much more fascinating to *do* mathematics than reading textbooks. The proof of relaxation generalized that found by Ludwig Boltzmann in 1872. I read Boltzmann's proof in a textbook of statistical mechanism fifteen years ago. Then, to make it self-consistent, I wrote an introduction to relative entropy, which was the starting point of the proof. Surprisingly, by appending locality into the axioms of information, relative entropy would be the unique possibility. The next surprise came from introducing smooth structure to stochastic system. I found the cut-off by a series of complex calculation. Later on in the winter, I found a way of expanding a function by a series of generalized functions. It led to a direct proof of Kramers-Moyal expansion. It also gave raise to a rigorous proof of path integral formulation for stochastic system, resulting in a new perspective to data-fitting.

After climbing to the top of a mountain, the horizon is broaden. Out of what we have explored, there arises more interesting questions. How does information propagate in a stochastic system? Why do the starlings in a flock, ants in a colony, and even trees in a forest behave as a whole like an individual organism? There are much more stories to be told, much more treasures to be sought, and much more beauties to be enjoyed. The new trip has been ahead.