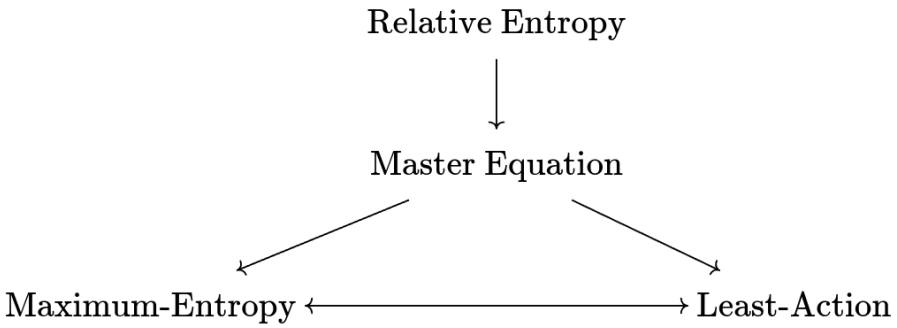

Figure 1. Scheme of sections. The solid arrow represents relation.

1 Relative Entropy

1.1 A Brief Review of Probability

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.

The set of all possible values of a random variable is called the **alphabet**.[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).[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 $p_Y(y) := \int_{\mathcal{Z}} \mathrm{d}z p(y,z)$ and $p_Z(z) := \int_{\mathcal{Y}} \mathrm{d}y p(y,z)$ if X is of continuum, and the same for mass function. We **marginalize** Z so as to get P_Y.

1.2 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

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

1. Some 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 book *Information Theory, Inference, and Learning Algorithms*, section 2.1. Link to free PDF: https://www.inference.org.uk/itprnn/book.pdf).

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.

where k is a positive constant. Interestingly, this expression is unique given some plausible conditions, 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 condition is also called the additivity of information. For two independent variables X and Y with distributions P and Q respectively, the third condition indicates that the total information of $H(PQ)$ is $H(P)+H(Q)$. But, the third condition 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 conditions.[3] This conditional entropy is, argued by others, quite strong and not sufficiently natural. The problem is that this stronger condition is essential for Shannon entropy to arise. Otherwise, there will be other entropy definitions that satisfy all the conditions, where the third involves only independent random variables, such as Rényi entropy.[4]

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

1.3 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 extends the conditions as

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 condition employs the locality of H, which is thought as natural since H has been a functional. The second condition indicates that H vanishes only when there is no surprise caused by P (thus $P = Q$). It is a little like the second condition for Shannon entropy. The third condition, 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.

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.

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.

1.4 Relative Entropy is the Unique Solution to the Conditions

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

Based on the first condition, 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 condition,

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

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 condition. 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 condition 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 $\partial_1 h(ax,by)\,x=\partial_1 h(a,b)$ and $\partial_2 h(ax,by)\,y=\partial_2 h(a,b)$. Since $\partial_1 h(a,a)+\partial_2 h(a,a)=(\mathrm{d}/\mathrm{d}a)\,h(a,a)=0$, we get $\partial_1 h(ax,ay)\,x+\partial_2 h(ax,ay)\,y=0$. Letting $a=1$, it becomes a first order partial differential equation $\partial_1 h(x,y)\,x+\partial_2 h(x,y)\,y=0$, 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 $\partial_1 h(x,1)=\partial_1 h(1,1)/x$, which has a unique solution that $h(x,1)=\partial_1 h(1,1)\ln(x)+C$, where C is a constant. Combined with $h(x,y)=h(x/y,1)$, we finally arrive at $h(x,y)=\partial_1 h(1,1)\ln(x/y)+C$. To determine the $\partial_1 h(1,1)$ and C, we use the second condition $\partial_1 h(1,1)\int\mathrm{d}x p(x)\ln(p(x)/q(x))+C>0$ when $p\neq q$ and $\partial_1 h(1,1)\int\mathrm{d}x p(x)\ln(p(x)/p(x))+C=0$. The second equation results in $C=0$. By Jensen's inequality, the integral $\int\mathrm{d}x p(x)\ln(p(x)/q(x))$ is non-negative, thus from the first equation, $\partial_1 h(1,1)>0$. Up to now, all things about h have been settled. We conclude that there is a unique expression that satisfies all the three conditions, 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 conditions when locality is absent.

In the end, we examine the two issues appeared in Shannon entropy (section 1.3). 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.

2 Master Equation, Detailed Balance, and Relative Entropy

2.1 Conventions in This Section

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 section 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 1.3), while the inverse will be safe and direct as long as any smooth structure of X is not employed throughout the discussion.

2.2 Master Equation Describes the Evolution of Markov Process

Without losing generality, consider a pile of sand on a desk. The desk has been fenced in so that the sands will not flow out of the desk. Imagine that these sands are magic, having free will to move on the desk. The distribution of sands changes with time. In the language of probability, the density of sands at position x of the desk is described by a time-dependent density function $p(x,t)$, where the total mass of the sands on the desk is normalized to 1, and the position on the desk characterizes the alphabet $\mathcal{X}$.

Let $q_{t\to t'}(y|x)$ denote the *portion* of density at position x that transits to position y, from t to t'. Then, the transited density will be $q_{t\to t'}(y|x)\,p(x,t)$. 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 imagine the sands transit from position x to x (stay on x), which is $q_{t\to t'}(x|x)$. Now, every sand at position x has transited during $t\to t'$, and the total portion shall be 100%, which means

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

As portion, $q_{t\to t'}$ cannot be negative, thus $q_{t\to t'}(x|y)\geqslant 0$ for each x and y 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 on density at position x. The difference is caused by the density transited from x, which is $\int_{\mathcal{X}}\mathrm{d}y\,q_{t\to t'}(y|x)\,p(x,t)$, and that transited to x, which is $\int_{\mathcal{X}}\mathrm{d}y q_{t\to t'}(x|y)p(y,t)$. Thus, we have

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

By inserting equation (1), we find

$$p(x,t') = \int_{\mathcal{X}} \mathrm{d}y q_{t\to t'}(x|y)p(y,t), \tag{2}$$

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

$$q_{t\to t}(x|y) = \delta(x-y).$$

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

$$\frac{\partial p}{\partial t}(x,t) = \int_{\mathcal{X}} \mathrm{d}y r_t(x,y)p(y,t),$$

where $r_t(x,y) := \lim_{t'\to t}(\partial q_{t\to t'}/\partial t')(x|y)$, called **transition rate**. It 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 interval $\Delta t := t' - t$, thus $q_{\Delta t}$. In this case, transition rate r is time-independent, so the master equation becomes

$$\frac{\partial p}{\partial t}(x,t) = \int_{\mathcal{X}} \mathrm{d}y\, r(x,y) p(y,t). \tag{3}$$

Since we only deal with Markovian stochastic process throughout this note, when referring to master equation, we mean equation 3. And to discrete time master equation, equation 4:

$$p(x,t+\Delta t) = \int_{\mathcal{X}} \mathrm{d}y\, q_{\Delta t}(x,y) p(y,t). \tag{4}$$

Before finishing this section, we discuss the demanded conditions for transition rate. The normalization of transition density 1 implies that $\int_{\mathcal{X}} \mathrm{d}x\, r(x,y) = 0$. This can be seen by Taylor expanding $q_{\Delta t}$ by Δt, as $q_{\Delta t}(x|y) = \delta(x-y) + r(x,y)\,\Delta t + o(\Delta t)$, where we have inserted $q_0(x|y) = \delta(x-y)$ 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 1.1), the equation 2 must conserve this positivity. We are to show that this is guaranteed by the master equation itself, without any extra condition demanded for the transition rate. It is convenient to use discrete notations, thus replace $x \to i$, $y \to j$, and $\int \to \sum$. The master equation turns to be $(\mathrm{d}p_i/\mathrm{d}t)(t) = \sum_j r_{ij}\, p_j(t)$. Notice that it becomes 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 $r_{ii} p_i(t) + \sum_{j:j\neq i} r_{ij} p_j(t)$, and the worst situation is that $r_{ij} = 0$ for each $j \neq i$ and $r_{ii} < 0$. In this case, the master equation reduces to $(\mathrm{d}p_i/\mathrm{d}t)(t) = r_{ii} p_i(t)$, which has the solution $p_i(t) = p_i(0)\exp(r_{ii} t)$. It implies that $p_i(t) > 0$ as long as $p_i(0) > 0$, 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 $\int_{\mathcal{X}} \mathrm{d}x\, r(x,y) = 0$.

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 (finite) from the local (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 3 to calculate $\partial^n p/\partial t^n$ for any n. Indeed, for $n = 2$,

$$\begin{aligned}
&\frac{\partial^2 p}{\partial t^2}(z,t) \\
\{\text{insert equation 3}\} &= \frac{\partial}{\partial t}\int_{\mathcal{X}} \mathrm{d}y\, r(z,y)\, p(y,t) \\
\{\text{exchange limits}\} &= \int_{\mathcal{X}} \mathrm{d}y\, r(z,y)\, \frac{\partial p}{\partial t}(y,t) \\
\{\text{insert equation 3}\} &= \int_{\mathcal{X}} \mathrm{d}y\, r(z,y) \int_{\mathcal{X}} \mathrm{d}x\, r(y,x)\, p(x,t) \\
\{\text{rearrange}\} &= \int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}y\, r(z,y)\, r(y,x)\, p(x,t).
\end{aligned}$$

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

$$\frac{\partial^{n+1}p}{\partial t^{n+1}}(z,t)=\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y_1\cdots\int_{\mathcal{X}}\mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1})\cdots r(y_1,x)\, p(x,t).$$

Notice the pattern: a sequence of r and a rightmost $p(x,t)$. The reason for this pattern to arise is that $q_{\Delta t}$, thus r, is independent of t: a Markovian property.

Also, Taylor expand the both sides of equation 4 by Δt gives, at $(\Delta t)^{n+1}$ order,

$$\frac{\partial^{n+1}p}{\partial t^{n+1}}(z,t)=\int_{\mathcal{X}}\mathrm{d}x\lim_{\Delta t\to 0}\frac{\partial^{n+1}q_{\Delta t}}{\partial(\Delta t)^{n+1}}(z|x)p(x,t).$$

So, we arrive at

$$\int_{\mathcal{X}}\mathrm{d}x\left[\lim_{\Delta t\to 0}\frac{\partial^{n+1}q_{\Delta t}}{\partial(\Delta t)^{n+1}}(z|x)-\int_{\mathcal{X}}\mathrm{d}y_1\cdots\int_{\mathcal{X}}\mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1})\cdots r(y_1,x)\right]p(x,t)=0,$$

which holds for all $p(x,t)$, thus

$$\lim_{\Delta t\to 0}\frac{\partial^{n+1}q_{\Delta t}}{\partial(\Delta t)^{n+1}}(z|x)=\int_{\mathcal{X}}\mathrm{d}y_1\cdots\int_{\mathcal{X}}\mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1})\cdots r(y_1,x),$$

or say[5]

$$\begin{aligned}q_{\Delta t}(z|x)=\;&\delta(z-x)\\+\;&(\Delta t)\, r(z,x)\\+\;&\frac{(\Delta t)^2}{2!}\int_{\mathcal{X}}\mathrm{d}y\, r(z,y)\, r(y,x)\\+\;&\cdots\\+\;&\frac{(\Delta t)^{n+1}}{(n+1)!}\int_{\mathcal{X}}\mathrm{d}y_1\cdots\int_{\mathcal{X}}\mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1})\cdots r(y_1,x)\\+\;&\cdots.\end{aligned}\tag{5}$$

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 infinitesimal time-interval transition density.*

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

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

also leads to equation 3, 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.

---

5. 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.

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

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

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

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

Thus,

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

Since $p(x,t)$ can be arbitrary, we arrive at

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

This provides an addition restriction to the transition density. Indeed, not every transition density, as a function of time interval Δt, can satisfy this relation.

2.4 Detailed Balance Provides Stationary Distribution

Let Π a stationary solution of master equation 3. Then, its density function π satisfies $\int_{\mathcal{X}} \mathrm{d}y r(x, y)\, \pi(y) = 0$. Since we have demanded that $\int_{\mathcal{X}} \mathrm{d}y r(y,x) = 0$, the stationary master equation can be re-written as

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

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

$$r(x,y)\, \pi(y) = r(y,x)\pi(x), \tag{6}$$

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

Interestingly, for a transition rate r that satisfies detailed balance condition 6, the transition density $q_{\Delta t}$ generated by r using equation 5 satisfies a similar relation

$$q_{\Delta t}(x|y)\, \pi(y) = q_{\Delta t}(y|x)\pi(x). \tag{7}$$

To see this, consider the third line in equation (5), where the main factor is

$$\begin{aligned} q_{\Delta t}(z|x)\, \pi(x) &\supset \int \mathrm{d}y r(z,y)\, r(y,x)\, \pi(x) \\ \{r(y,x)\, \pi(x) = \pi(y)\, r(x,y)\} &= \int \mathrm{d}y r(z,y)\, \pi(y)\, r(x,y) \\ \{r(z,y)\, \pi(y) = \pi(z)\, r(x,y)\} &= \int \mathrm{d}y \pi(z)\, r(x,y)\, r(y,z) \\ &= \pi(z) \int \mathrm{d}y r(x,y)\, r(y,z) \\ &\subset q_{\Delta t}(x|z)\, \pi(z) \end{aligned}$$

Following the same steps, we can show that all terms in equation 5 share the same relation, indicating $q_{\Delta t}(z|x)\, \pi(x) = q_{\Delta t}(x|z)\, \pi(z)$.

2.5 Detailed Balance Condition and Connectivity Monotonically Reduce 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 $x\in\mathcal{X}$, we have defined the relative entropy between them, as

$$H(P(t),\Pi)=\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)\ln\frac{p(x,t)}{\pi(x)}. \tag{8}$$

It describes the uncertainty (surprise) caused by $P(t)$ when prior knowledge is given by Π. It is a plausible generalization of Shannon entropy to continuous random variables.

We can calculate the time-derivative of relative entropy by master equation 3. Generally, the time-derivative of relative entropy has no interesting property. But, if the π is more than stationary but satisfying a stronger condition: detailed balance, then $\mathrm{d}H(P(t),\Pi)/\mathrm{d}t$ will have a regular form[6]

$$\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)=-\frac{1}{2}\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\,r(x,y)\,\pi(x)\left(\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right)\left(\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right). \tag{9}$$

6. 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 3, as

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

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

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

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

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

By adding the two previous results together, we find

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

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 $r(x,y)\,\pi(y)$ (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 $r(y_{i+1},y_i)$ and $r(y_i,y_{i+1})$ 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 $p(x,t)/\pi(x)=p(z,t)/\pi(z)$.[7] So, by repeating the previous discussion on 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 densities 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 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 6 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 3, $P(t)$ 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 3 that tends to equilibrium (without which the simulation will not terminate)? Using the metaphor of sands (see section 2.2), we simulate each sand, but replace its free will by a transition probability. Explicitly, we initialize the sands (that is, their positions) randomly. Then iteratively update the position of each sand. In each iteration, a sand jumps from position x to position y with the probability $q_{\Delta t}(y|x)\approx\delta(y-x)+r(y,x)\,\Delta t$ where Δt is sufficiently small. Not every jump is valid. On one hand, we have to ensure that

7. 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 $z=y_n$.

computer has a sampler that makes random sampling for $q_{\Delta t}(y|x)$. On the other hand, to ensure the termination, the transition rate r, together with the density function π, shall satisfy the detailed balance condition 6. (Section 2.7 will provide a method that constructs such a transition rate from the density function.) 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.[8]

Like the Euler method in solving dynamical system, however, a finite time step results in a 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

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

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 4 instead, and detailed balance condition 7 for transition density), the second line reduces to

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

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 $\Delta t\to 0$.

8. 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, whose density function is the π. By central limit theorem (briefly, the mean of i.i.d. random variables $X_1,\dots,X_N$ with mean $\mathbb{E}[X_i]=0$ and variance $\mathrm{Var}[X_i]=\sigma^2$ for some σ, has standard derivation $\sigma/\sqrt{N}$ when N is large enough), 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 1.1, 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 ΔH_2. Taylor expanding ΔH_1 by Δt gives[9]

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

where, by master equation 3, $(\partial/\partial t)\ln p(x,t) = \int_{\mathcal{X}} \mathrm{d}x\, r(x,y)\, p(y,t)/p(x,t)$. For ΔH_2, we insert equation 9 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 (represented by 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 function has a greater slope and we can safely employ a relatively larger learning rate to speed up the decreasing of loss. But, we have to tune the learning rate to be smaller and smaller during the training, in which the slope of loss function 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.)

---

9. 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 a stationary distribution Π, and an auxiliary transition rate γ, ensuring that $\gamma(x,y)>0$ for each x and y in alphabet $\mathcal{X}$ such that $x\neq y$, the transition rate r is given by

$$r(x,y)=\min\left(1,\frac{\gamma(y,x)\,\pi(x)}{\gamma(x,y)\,\pi(y)}\right)\gamma(x,y). \tag{10}$$

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 the detailed balance condition 6. Directly,

$$\begin{aligned}
& r(x,y)\,\pi(y)\\
\{\text{definition of } r\} &= \min\left(1,\frac{\gamma(y,x)\,\pi(x)}{\gamma(x,y)\,\pi(y)}\right)\gamma(x,y)\,\pi(y)\\
\{\text{property of min}\} &= \min\left(\gamma(x,y)\,\pi(y),\gamma(y,x)\,\pi(x)\right)\\
\{\text{property of min}\} &= \min\left(\frac{\gamma(x,y)\,\pi(y)}{\gamma(y,x)\,\pi(x)},1\right)\gamma(y,x)\,\pi(x)\\
\{\text{definition of } r\} &= r(y,x)\,\pi(x).
\end{aligned}$$

Thus detailed balance condition holds. So, theorem 1 states that, *evolved by the master equation 3, 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

$$q(x|y):=\min\left(1,\frac{g(y|x)\,\pi(x)}{g(x|y)\,\pi(y)}\right)g(x|y), \tag{11}$$

where g is positive-definite on $\mathcal{X}$. Notice that, in equation 11 there is no extra time parameter like the $q_{\Delta t}(x|y)$ 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 $g(x|y)$ first proposes a transition from y to x. (In numerical simulation, we have to ensure that computer has a sampler for sampling an x from the conditional probability $g(x|y)$.) 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 x, otherwise stay on y. Altogether, we get a conditional probability jumping from y to x, the $q(x|y)$.

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 expand $g_{\Delta t}$ at $\Delta t\to 0$ as $g_{\Delta t}(x|y)=\delta(x-y)+\gamma(x,y)\,\Delta t+o(\Delta t)$, then we will find $q_{\Delta t}(x|y)=\delta(x-y)+r(x,y)\,\Delta t+o(\Delta t)$. Indeed, when $x=y$, we have $q_{\Delta t}(x|x)=g_{\Delta t}(x,x)$. And when $x\neq y$, $\delta(x-y)=0$, we find

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

Altogether, for each $x,y\in\mathcal{X}$, we find $q_{\Delta t}(x|y)=\delta(x-y)+r(x,y)\,\Delta t+o(\Delta t)$. *In practice, we use the Metropolis-Hastings algorithm 11 to numerically simulate master equation 3.* But, based on the discussion in section 2.6, the Δt in $g_{\Delta t}$ shall be properly bounded to be small (or equivalently speaking, g shall be "principal diagonal") so as to ensure the relaxation $P(t)\to\Pi$.

2.8 * Existence of Stationary Density Function

Given a transition rate, we wonder if there exists a density function such that detailed balance condition 6 holds. Actually, equation 6 *defines* a density function. For example, if both $r(x,y)$ and $r(y,x)$ are not zero, we can construct $\pi(y)$ by given $\pi(x)$ as $\pi(y)=\pi(x)\,r(y,x)/r(x,y)$. Generally, if x and y are connected, then there is a path $P:=(p_0,\dots,p_n)$ from x to y with $p_0=x$ and $p_n=y$ (path and connectivity are defined in section 2.5), and define

$$\begin{aligned}\pi(p_1)&:=\pi(p_0)\,r(p_1,p_0)/r(p_0,p_1)\\ \pi(p_2)&:=\pi(p_1)\,r(p_2,p_1)/r(p_1,p_2)\\ &\cdots\\ \pi(p_n)&:=\pi(p_{n-1})\,r(p_n,p_{n-1})/r(p_{n-1},p_n).\end{aligned}$$

Thus, $\pi(y)$ (the $\pi(p_n)$) is constructed out of $\pi(x)$ (the $\pi(p_0)$). Let $\rho(x,y):=\ln r(x,y)-\ln r(y,x)$, it becomes

$$\ln\pi(y)=\ln\pi(x)+\sum_{i=0}^{n-1}\rho(p_{i+1},p_i),$$

or in continuous format,

$$\ln\pi(y)=\ln\pi(x)+\int_P \mathrm{d}s\,\rho(s), \tag{12}$$

where $\rho(s)$ is short for $\rho(p_{s+1},p_s)$ along the path P. In this way, given $x_0\in\mathcal{X}$, we define any $x\in\mathcal{X}$ that is connected to x_0 by $\ln\pi(x):=\ln\pi(x_0)+\int_P \mathrm{d}s\,\rho(s)$. And $\pi(x_0)$ is determined by the normalization of π.

But, there can be multiple paths from x to y which are connected in $\mathcal{X}$. For example, consider two paths P and P', then we have $\int_P \mathrm{d}s\,\rho(s)=\int_{P'}\mathrm{d}s\,\rho(s)$. Generally, if C is a **circle** which is a path starting at an element $x\in\mathcal{X}$ and finally end at x (but not simply standing at x), then

$$\oint_C \mathrm{d}s\,\rho(s)=0. \tag{13}$$

It means every path along two connected elements in $\mathcal{X}$ is equivalent. If the condition 13 holds, we can simplify the notation in equation 12 by

$$\ln\pi(y)=\ln\pi(x)+\int_x^y \mathrm{d}s\,\rho(s),$$

where $\int_x^y$ indicates any path from x to y (if x and y are connected).

Condition 13 implies that the previous construction does define a π that holds the detailed balance condition. Given $x,y\in\mathcal{X}$, we have $\ln\pi(x)=\ln\pi(x_0)+\int_{x_0}^x \mathrm{d}s\,\rho(s)$ and $\ln\pi(y)=\ln\pi(x_0)+\int_{x_0}^y \mathrm{d}s\,\rho(s)$. If x and y are connected, then, by condition 13, $\rho(y,x)=\int_x^{x_0}\mathrm{d}s\,\rho(s)+\int_{x_0}^y \mathrm{d}s\,\rho(s)$ (the $\rho(y,x)$ indicates the path (x,y), "jumping" directly from x to y), thus $\ln\pi(y)=\ln\pi(x)+\rho(y,x)$, which is just the detailed balance condition 6. And if x and y are not connected, then both $r(x,y)$ and $r(y,x)$ shall vanish (recall the requirements of transition rate in section 2.2: if $r(x,y)=0$, then $r(y,x)=0$), and detailed balance condition holds naturally.

So, condition 13 is *essential and sufficient for the existence of π that holds the detailed balance condition 6*. If $\mathcal{X}$ is a simply connected smooth manifold, then using Stokes's theorem, we have $\nabla\times\rho=0$ on $\mathcal{X}$. But, generally $\mathcal{X}$ is neither simply connected nor smooth, but involving independent subsystems and discrete. In these cases, condition 13 becomes very complicated.

In many applications, we consider the inverse question: given a density function, if there exists a transition rate such that detailed balance condition holds. This inverse problem is much easier, and a proper transition rate can be constructed out of the density function (such as in Metropolis-Hastings algorithm).

3 Kramers-Moyal Expansion and Langevin Process

We follow the discussion in section 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 section 2 hold in this section, but the inverse is not guaranteed.

3.1 Conventions in This Section

Follow the conventions in section 2. In addition, we employ the Einstein's convention of summation. 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 $\sum_\alpha \sum_\beta \partial_\alpha A^\alpha_\beta\, x^\beta$.

3.2 Spatial Expansion of Master Equation Gives Kramers-Moyal Expansion

Let the alphabet $\mathcal{X} = \mathbb{R}^n$ for some integer $n \geqslant 1$, which has sufficient connectivity. In addition, suppose that the density function $p(x,t)$ of a time-dependent distribution $P(t)$ and the transition rate $r(x,y)$ are smooth on x and y. In this section, we investigate the direct results of spatial smoothness.

Now, the master equation 3 becomes

$$\frac{\partial p}{\partial t}(x,t) = \int_{\mathbb{R}^n} \mathrm{d}y r(x,y)\, p(y,t).$$

The spatial smoothness indicates that we can Taylor expand the right hand side to arbitrary order. The quantity that is used to perform the Taylor expansion neither x nor y since they are equally weighted, but their difference, $\epsilon := x - y$. If we replace the y in the right hand side with $x - \epsilon$, that is, $\int_{\mathbb{R}^n} \mathrm{d}y r(x,y)\, p(y,t) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x, x-\epsilon)\, p(x-\epsilon, t)$, and directly Taylor expand by ϵ, then we will get the leading term $\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x,x)\, p(x,t)$, the result of which is unknown. What we have known is $\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x+\epsilon, x)\, p(x,t)$ which is zero because of the "normalization" of transition density. So, we expect to Taylor expand by ϵ that which results in a leading term $\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x+\epsilon, x)\, p(x,t)$. To do this, we need a little magic.

First of all, we have the identity

$$\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x, x-\epsilon)\, p(x-\epsilon, t) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon r((x-\epsilon)+\epsilon, x-\epsilon)\, p(x-\epsilon, t).$$

Next, we perform the magic. We first define $\omega(x,\epsilon) := r(x+\epsilon, x)$, which the factor we want to obtain in the leading term. Then, the integral turns to be $\int_{\mathbb{R}^n} \mathrm{d}\epsilon \omega(x-\epsilon, \epsilon)\, p(x-\epsilon, t)$. The key is Taylor expanding by the ϵ *in the first argument* of $\omega(x-\epsilon,\epsilon)$ in addition to that in $p(x-\epsilon,t)$. So, it becomes

$$\int_{\mathbb{R}^n} \mathrm{d}\epsilon \omega(x,\epsilon)\, p(x,t) + \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k}) \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \left(\frac{\partial}{\partial x^{\alpha_1}} \cdots \frac{\partial}{\partial x^{\alpha_k}} \right) [\omega(x,\epsilon)\, p(x,t)].$$

The leading term (the first one) vanishes, as expected. With some re-arrangement to the second term, and plugging it back to the right hand side of master equation, we find

$$\frac{\partial p}{\partial t}(x,t) = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \left(\frac{\partial}{\partial x^{\alpha_1}} \cdots \frac{\partial}{\partial x^{\alpha_k}} \right) \left[p(x,t) \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, \omega(x,\epsilon) \right].$$

The integral $\int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, \omega(x,\epsilon)$ in the $[\cdots]$ factor has an intuitive meaning. Remind of $\omega(x,\epsilon) = r(x+\epsilon, x)$ and $q_{\Delta t}(x+\epsilon|x) = \delta(\epsilon) + r(x+\epsilon, x)\, \Delta t + o(\Delta t)$, we have

$$\int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, q_{\Delta t}(x+\epsilon|x) = \Delta t \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, \omega(x,\epsilon) + o(\Delta t).$$

So, $\Delta t \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, \omega(x,\epsilon)$ is recognized as an approximation of the k-order correlation of ϵ sampled from transition density $q_{\Delta t}(x+\epsilon|x)$ (regarding $q_{\Delta t}(x+\epsilon|x)$ as an x-dependent distribution $Q_{\Delta t}(x)$ that samples ϵ). We denote it by (K for the leading consonant of "correlation")

$$K^{\alpha^1 \cdots \alpha^k}(x) := \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k})\, \omega(x,\epsilon).$$

Finally, we arrive at

$$\frac{\partial p}{\partial t}(x,t)=\sum_{k=1}^{+\infty}\frac{(-1)^k}{k!}\left(\frac{\partial}{\partial x^{\alpha_1}}\cdots\frac{\partial}{\partial x^{\alpha_k}}\right)[K^{\alpha^1\cdots\alpha^k}(x)\,p(x,t)]. \tag{14}$$

This Taylor expansion of master equation is called the **Kramers–Moyal expansion**.

3.3 Transition Density of Langevin Process Is Approximately Gaussian

Spatial connectivity enables us to investigate the most general distribution of continuous random variable, the normal distribution. Given $\mu\colon\mathbb{R}^n\to\mathbb{R}^n$ and $\Sigma\colon\mathbb{R}^n\to\mathbb{R}^{n\times n}$, which is positive definite and symmetric, let $q_{\Delta t}(x'|x)$ the normal distribution of $x'-x$ with mean value $\mu(x)\,\Delta t$ and variance $2\Sigma(x)\Delta t$. That is,

$$q_{\Delta t}(x+\epsilon|x)=\frac{1}{\sqrt{(4\pi\Delta t)^n\det\Sigma(x)}}\exp\left(-\frac{1}{4\Delta t}\,[\Sigma^{-1}(x)]_{\alpha\beta}\,(\epsilon^\alpha-\mu^\alpha(x)\,\Delta t)\,(\epsilon^\beta-\mu^\beta(x)\,\Delta t)\right).$$

When Δt is sufficiently small, $q_{\Delta t}$ can be approximately regarded as a transition density (section 2.3). The corresponding Markov process is called **Langevin dynamics** or **Langevin process**.

In many textures, Langevin process is written by a stochastic differential equation (again, we use capital letters for random variables)

$$\mathrm{d}X^\alpha=\mu^\alpha(x)\mathrm{d}t+\mathrm{d}W^\alpha(x),$$

where $\mathrm{d}W^\alpha(x)$, called **Wiener process**, is a random variable obeying the normal distribution with zero mean and variance $2\,\Sigma(x)\,\mathrm{d}t$. This stochastic differential equation is an formal equivalent of $q_{\Delta t}(x+\epsilon|x)$ when $\Delta t\to\mathrm{d}t$ and $\epsilon\to\mathrm{d}x$.

3.4 Transition Rate of Langevin Process Is a Generalized Function

In this section, we calculate the the transition rate of Langevin process from transition density. The Δt appears in many places in transition density, and directly Taylor expanding $q_{\Delta t}$ by Δt is very hard. Instead, we employ an arbitrary test function $f\in S(\mathbb{R}^n,\mathbb{R})$,[10] and Taylor expand f by its variable

$$\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)\,f(\epsilon)=\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)\left[\,f(0)+\epsilon^\alpha\,\partial_\alpha f(0)+\frac{1}{2}\epsilon^\alpha\,\epsilon^\beta\,\partial_\alpha\partial_\beta f(0)+\cdots\right]$$

These Gaussian integrals result in

$$\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)\,\epsilon^\alpha=\mu^\alpha(x)\,\Delta t$$

and (recall the relation between covariance and mean, $\mathrm{Cov}(X,Y)=\mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y]$)

$$\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)\,\epsilon^\alpha\,\epsilon^\beta=\Sigma^{\alpha\beta}(x)\,\Delta t+\mu^\alpha(x)\,\mu^\beta(x)\,\Delta t^2=\Sigma^{\alpha\beta}(x)\,\Delta t+o(\Delta t).$$

Altogether,

$$\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)\,f(\epsilon)=f(0)+\Delta t[\,\mu^\alpha(x)\,\partial_\alpha f(0)+\Sigma^{\alpha\beta}(x)\,\partial_\alpha\partial_\beta f(0)]+o(\Delta t),$$

as $\Delta t\to 0$ (for example, $\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,q_{\Delta t}(x+\epsilon|x)[\,\epsilon^\alpha\epsilon^\beta\epsilon^\gamma\epsilon^\delta\,\partial_\alpha\partial_\beta\partial_\gamma\partial_\delta f(0)]=\mathcal{O}(\Delta t^2)=o(\Delta t)$). On the other hand, if we Taylor expand $q_{\Delta t}$ by Δt as $q_{\Delta t}(x+\epsilon|x)=\delta(\epsilon)+r(x+\epsilon,x)\,\Delta t+o(\Delta t)$, where r is the transition rate, then we will get

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

10. The S represents Schwarts space, which is a functional space in which any function $f\colon X\to Y$ is smooth and rapidly falls to zero in the region far from origin. For example, Gaussian function (the density function of normal distribution) is in $S(\mathbb{R},\mathbb{R})$.

From the terms proportional to Δt, we recognize

$$\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x+\epsilon, x)\, f(\epsilon) = \mu^\alpha(x)\, \partial_\alpha f(0) + \Sigma^{\alpha\beta}(x)\, \partial_\alpha \partial_\beta f(0).$$

Noticing the integration by parts[11]

$$-\int_{\mathbb{R}^n} \mathrm{d}\epsilon \mu^\alpha(x)\, \partial_\alpha \delta(\epsilon)\, f(\epsilon) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon \mu^\alpha(x)\, \delta(\epsilon)\, \partial_\alpha f(\epsilon) = \mu^\alpha(x)\, \partial_\alpha f(0),$$

and

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

we get

$$r(x+\epsilon, x) = -\mu^\alpha(x)\, \partial_\alpha \delta(\epsilon) + \Sigma^{\alpha\beta}(x)\, \partial_\alpha \partial_\beta \delta(\epsilon). \tag{15}$$

Because of the Dirac's δ-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 $f \in S(\mathbb{R}^n, \mathbb{R}^n)$, and calculate $\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(-x)\, f^\alpha(x)$. 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

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

Then, integration by parts gives $-\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha [\delta(-x)]\, f^\alpha(x) = \int_{\mathbb{R}^n} \mathrm{d}x\, \delta(-x)\, \partial_\alpha f^\alpha(x)$. After inserting the relation $\delta(x) = \delta(-x)$, we arrive at $\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(-x)\, f^\alpha(x) = \partial_\alpha f^\alpha(0)$. On the other hand, we have, by integration by parts, $-\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(x)\, f^\alpha(x) = \int_{\mathbb{R}^n} \mathrm{d}x\, \delta(x)\, \partial_\alpha f^\alpha(x) = \partial_\alpha f^\alpha(0)$. Altogether, we find $\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(-x)\, f^\alpha(x) = -\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(x)\, f^\alpha(x)$, for any $f \in S(\mathbb{R}^n, \mathbb{R}^n)$. 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*.[12] These conclusions are to be used in section 3.7.

---

11. High-dimensional integration by parts employs Stokes theorem. Consider the integral $\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \varphi(x)\, v^\alpha(x)$ with smooth scalar function $\varphi: \mathbb{R}^n \to \mathbb{R}$ and vector field $v: \mathbb{R}^n \to \mathbb{R}^n$. We have identity

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

The first integrand in the right hand side is a divergence. Using Stokes theorem, it becomes

$$\int_{\partial \mathbb{R}^n} \mathrm{d}S_\alpha\, [\varphi(x)\, v^\alpha(x)],$$

where $\partial \mathbb{R}^n$ is the "boundary" of $\mathbb{R}^n$. If φ or v is in Schwarts space, then this term vanishes, and the integral results in

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

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

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

Then integration by parts gives

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

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

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

So,

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

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

3.5 Master Equation of Langevin Process Is Fokker-Planck Equation

After discussing transition rate, we turn to the master equation of Langevin process. Since Langevin process applies to continuous random variable, we can use Kramers-Moyal expansion to evaluate its master equation. Directly, we have $K^\alpha(x)=\mu^\alpha(x)$, $K^{\alpha\beta}(x)=2\Sigma^{\alpha\beta}(x)$, and higher orders are all vanishing (K was defined in section 3.2). For example, the integral $\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,(\epsilon^\alpha\,\epsilon^\beta\,\epsilon^\gamma)\,q_{\Delta t}(x+\epsilon|x)=o(\Delta t)$. By relation $\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,(\epsilon^\alpha\,\epsilon^\beta\,\epsilon^\gamma)\,q_{\Delta t}(x+\epsilon|x)=\Delta t\,K^{\alpha\beta\gamma}(x)+o(\Delta t)$ (derived in section 3.2), we find $K^{\alpha\beta\gamma}(x)=0$. Thus, Kramers-Moyal expansion 14 reads

$$\frac{\partial p}{\partial t}(x,t)=-\partial_\alpha(\mu^\alpha(x)\,p(x,t))+\partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,p(x,t)). \tag{16}$$

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.

3.6 Stationary Solution of Fokker-Planck Equation

Fokker-Planck equation 16 has stationary solution Π which satisfies (since there is only one variable x, we use ∂ instead of ∇)

$$-\partial_\alpha(\mu^\alpha(x)\,\pi(x))+\partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x))=0,$$

which means

$$\mu^\alpha(x)\,\pi(x)=\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x))+\nu^\alpha(x), \tag{17}$$

where $\nu\colon\mathbb{R}^n\to\mathbb{R}^n$ 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 $\mathbb{R}^n$, we find that there is not net flux flowing out of anywhere in $\mathbb{R}^n$. Otherwise, suppose there is $x\in\mathbb{R}^n$ 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**.

3.7 Detailed Balance Condition for 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 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 15, we get, for the left hand side,

$$r(x+\epsilon,x)\,\pi(x)=-\mu^\alpha(x)\,\pi(x)\,\partial_\alpha\delta(\epsilon)+\Sigma^{\alpha\beta}(x)\,\pi(x)\,\partial_\alpha\partial_\beta\delta(\epsilon),$$

and, for the right hand side,

$$\begin{aligned}
&r(x,x+\epsilon)\,\pi(x+\epsilon)\\
=&\,r((x+\epsilon)-\epsilon,x+\epsilon)\,\pi(x+\epsilon)\\
=&-\mu^\alpha(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\delta(-\epsilon)+\Sigma^{\alpha\beta}(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\partial_\beta\delta(-\epsilon)\\
=&\,\mu^\alpha(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\delta(\epsilon)+\Sigma^{\alpha\beta}(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\partial_\beta\delta(\epsilon),
\end{aligned}$$

where in the last line, we use $\partial_\alpha\delta(-x)=-\partial_\alpha\delta(x)$ and $\partial_\alpha\partial_\beta\delta(-x)=\partial_\alpha\partial_\beta\delta(x)$.

As generalized functions, we are to examine these two expressions by using an arbitrary test function f. Thus, for the left hand side,

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

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

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

The right hand side is a little complicated,

$$\int_{\mathbb{R}^n} \mathrm{d}\epsilon r(x, x+\epsilon)\,\pi(x+\epsilon)\,f(\epsilon) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon \mu^\alpha(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\delta(\epsilon)\,f(\epsilon) + \int_{\mathbb{R}^n} \mathrm{d}\epsilon \Sigma^{\alpha\beta}(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\partial_\beta\delta(\epsilon)\,f(\epsilon).$$

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

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

By equaling $\int_{\mathbb{R}^n}\mathrm{d}\epsilon r(x+\epsilon, x)\,\pi(x)\,f(\epsilon)$ and $\int_{\mathbb{R}^n}\mathrm{d}\epsilon r(x, x+\epsilon)\,\pi(x+\epsilon)\,f(\epsilon)$, since f is arbitrary, we find, for the $f(0)$ terms,

$$-\partial_\alpha(\mu^\alpha(x)\,\pi(x)) + \partial_\alpha\partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)) = 0,$$

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

$$-\mu^\alpha(x)\,\pi(x) + \partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)) = 0.$$

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

$$\mu^\alpha(x)\,\pi(x) = \partial_\beta(\Sigma^{\alpha\beta}(x)\,\pi(x)). \tag{18}$$

Comparing with the stationary Fokker-Planck equation 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.

4 Least-Action Principle

In this section, we are to find a way of extracting dynamics (action or Lagrangian) from any raw data of any entity.

4.1 Conventions in This Section

Follow the conventions in section 3. 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, $\mathbb{R}^n$ for some n. A **path** is a function with single parameter $x\colon [t_i, t_f] \to \mathbb{R}^n$, 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, $\mathcal{P}(x_0, x_1) := \{x\colon [0,1] \to \mathbb{R}^n | x(0) = x_0, x(1) = x_1\}$ 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

$$S_{\mathrm{HO}}(x|x_0,x_1) = \frac{1}{2}\int_0^1 \mathrm{d}t\, [\dot{x}^2(t) - \omega^2 x^2(t)], \tag{19}$$

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

$$\frac{\partial S}{\partial x(i/T)}(x_\star|x_0,x_1) = 0, \tag{20}$$

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

Take harmonic oscillator as example. To discretize its action (equation 19), 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 19 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 $\ddot{x}_\star(t) + \omega^2 x_\star(t) = 0$.[13]

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

$$\frac{\partial S}{\partial x^\alpha}(x_\star) = 0, \tag{21}$$

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 21. 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, in statistical mechanics, as we will see in section TODO, the action is completely determined by the real world distribution (the correspondence of real world datum in statistical mechanics), 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, in which randomness is forbidden. There are, however, many phenomena in nature that have *intrinsic* randomness. For example, Langevin process $\mathrm{d}X = \mu(x)\,\mathrm{d}t + \mathrm{d}W$, which was originally used to describe molecular movement, has a stochastic term $\mathrm{d}W$ obeying a normal distribution with variance proportional to $\mathrm{d}t$. 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. Precisely, we use a distribution Q to describe real world phenomenon that has intrinsic randomness.

For any density function $q(x)$ and any $\beta > 0$, we can always define

$$S(x) := -(1/\beta) \ln q(x) + \mathrm{const}, \tag{22}$$

up to an arbitrary constant. Thus, $q(x) = \exp(-\beta S(x)) / Z$ where $Z := \int_{\mathcal{X}} \mathrm{d}x \exp(-\beta S(x))$. This S has some properties that can be analog to the action in classical mechanics. First, if $\mathcal{X} = \mathbb{R}^n$, then we find, by plugging in the definition of S,

$$\int_{\mathbb{R}^n} \mathrm{d}x\, q(x) \frac{\partial S}{\partial x^\alpha}(x) = -\beta^{-1} \int_{\mathbb{R}^n} \mathrm{d}x\, q(x) \frac{\partial}{\partial x^\alpha} \ln q(x) = -\beta^{-1} \int_{\mathbb{R}^n} \mathrm{d}x \frac{\partial}{\partial x^\alpha} q(x).$$

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. So, we conclude that

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

13. 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.

This is analog to equation 21, where the minimum $x_\star$ is replaced by the expectation $\mathbb{E}_Q$. Secondly, in the limit $\beta \to +\infty$ while fixing S, the distribution Q becomes so sharp that it only samples the $x_\star$ (recall section 1.1 that distribution has a sampler) 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 its density function q), without any redundancy. This is the direct implication that distribution involves more information than its most likely datum.

4.4 Example: How Far Will Information Propagate in a Stochastic System?

We are to construct a stochastic system to examine how far will information propagate from one side of the system to the other. For example, consider a 3-dimensional random variable $X = (X^1, X^2, X^3)$ in which there are interactions between X^1 and X^2, and between X^2 and X^3, but not (directly) between X^1 and X^3. It means that the action $S(x)$ that characterizes the distribution of X (see section 4.3) has the property that $\partial_1\partial_2 S(x)$ and $\partial_2\partial_3 S(x)$ are not always zero but $\partial_1\partial_3 S(x)$ is Even though, there is indirect interaction between X^1 and X^3, through X^2. But, this is not like deterministic system, such as a string, where information from one side can be safely propagated to the other side. Stochastic system has randomness. Information of X^1 is propagated to X^2 by the interaction between them. But, before propagating from X^2 to X^3, the information will be interrupted by the stochastic fluctuation of X^2. So, the information of X^1 shrinks when it arrives at X^3. It is how rumor spreads among people: rumor soon changes its face, becoming anything but itself. We wonder, if the propagation continues, along the chain of interaction, how far will it arrive before completely lost?

We are to examine this on an explicit stochastic system. The system shall be sufficiently simple so as to make analytical calculation. Consider the density function

$$q(x) \propto \exp\bigg(-\frac{\beta}{2} \int_{\mathbb{R}} \mathrm{d}t \, [\dot{x}^2(t) - \omega^2 \, x^2(t)] \bigg).$$

This action is not that of harmonic oscillator since the boundaries are not completely fixed (this is the key of simplification) and the time runs over $\mathbb{R}$. It turns out that discrete system is more complicated than this.

We follow the general strategy of dealing with integral that has derivatives in integrand: Fourier transform, by which we can decompose the interaction caused by derivative $\dot{x}(t)$ into independent variables. Denoting $\hat{x}(\zeta) \in \mathbb{C}$ as the Fourier coefficient of x, we have

$$\int_{\mathbb{R}} \mathrm{d}t \, \dot{x}^2(t) = \int_{\mathbb{R}} \mathrm{d}t \int_{\mathbb{R}} \mathrm{d}\zeta \int_{\mathbb{R}} \mathrm{d}\zeta' \, (-\zeta\zeta') \exp(\mathrm{i} \, (\zeta + \zeta') \, t) \, \hat{x}(\zeta) \, \hat{x}(\zeta') = \int_{\mathbb{R}} \mathrm{d}\zeta \, \zeta^2 \, \hat{x}(\zeta) \, \hat{x}(-\zeta),$$

where we have used $\int_{\mathbb{R}} \mathrm{d}t \exp(i \, k t) = \delta(k)$. Since $x(t)$ is real, we have $\hat{x}(-\zeta) = \overline{\hat{x}(\zeta)}$, and only the $\hat{x}(\zeta)$s with $\zeta \in [0, +\infty)$ are independent variables. Thus,

$$\int_{\mathbb{R}} \mathrm{d}t \, \dot{x}^2(t) = \int_{\mathbb{R}} \mathrm{d}\zeta \zeta^2 \, |\hat{x}(\zeta)|^2 = 2 \int_0^{+\infty} \mathrm{d}\zeta \zeta^2 \, |\hat{x}(\zeta)|^2.$$

The same, we have $\int_0^1 \mathrm{d}t \, x^2(t) = 2 \int_0^{+\infty} \mathrm{d}\zeta \, |\hat{x}(\zeta)|^2$. In addition, we decompose the complex number $\hat{x}(\zeta) = a(\zeta) + \mathrm{i} b(\zeta)$ with $a(\zeta), b(\zeta) \in \mathbb{R}$, thus $|\hat{x}(\zeta)|^2 = a^2(\zeta) + b^2(\zeta)$. Altogether, we find

$$q(a, b) \propto \exp\bigg(-\beta \int_0^{+\infty} \mathrm{d}\zeta \, (\zeta^2 + \omega^2) \, [a^2(\zeta) + b^2(\zeta)] \bigg).$$

Now, we get discrete (but infinitely many) variables that do not interact with each other.

After decomposing the variables, we are to compute how far will the information propagate along the time-axis. Pearson coefficient between $x(t)$ and $x(t')$ for $t, t' \in (0,1)$ is one quantity that characterizes our purpose. It is defined as $\mathrm{Cov}_Q(x(t),x(t'))/\sqrt{\mathrm{Var}_Q[x(t)]\,\mathrm{Var}_Q[x(t')]}$. Since $\mathrm{Var}_Q[x(t)] = \mathrm{Cov}_Q(x(t),x(t))$, all we need to do is calculating the covariance. Inserting the Fourier transform of $x(t)$, the bi-linearity of covariance gives

$$\mathrm{Cov}_Q(x(t),x(t')) = \int_{\mathbb{R}} \mathrm{d}\zeta \int_{\mathbb{R}} \mathrm{d}\zeta' \exp(\mathrm{i}\,\zeta t)\exp(\mathrm{i}\,\zeta' t')\,\mathrm{Cov}_Q(\hat{x}(\zeta),\hat{x}(\zeta')).$$

Then, we use the real variables $a(\zeta)$ and $b(\zeta)$ for which $\zeta \in [0,+\infty)$. This will separate the integral into four parts:

$$\begin{aligned}
&\mathrm{Cov}_Q(x(t),x(t')) \\
&= \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(\mathrm{i}\,\zeta t)\exp(\mathrm{i}\,\zeta' t')\,\mathrm{Cov}_Q(\hat{x}(\zeta),\hat{x}(\zeta')) \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(-\mathrm{i}\,\zeta t)\exp(\mathrm{i}\,\zeta' t')\,\mathrm{Cov}_Q(\hat{x}(-\zeta),\hat{x}(\zeta')) \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(\mathrm{i}\,\zeta t)\exp(-\mathrm{i}\,\zeta' t')\,\mathrm{Cov}_Q(\hat{x}(\zeta),\hat{x}(-\zeta')) \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(-\mathrm{i}\,\zeta t)\exp(-\mathrm{i}\,\zeta' t')\,\mathrm{Cov}_Q(\hat{x}(-\zeta),\hat{x}(-\zeta')).
\end{aligned}$$

Before replacing $\hat{x}(\zeta)$ by $a(\zeta)+\mathrm{i}b(\zeta)$, notice that $a(-\zeta)=a(\zeta)$ and $b(-\zeta)=-b(\zeta)$, as a result of $\hat{x}(-\zeta)=\overline{\hat{x}(\zeta)}$. Then we have, for example, $\mathrm{Cov}_Q(\hat{x}(-\zeta),\hat{x}(\zeta')) = \mathrm{Cov}_Q(a(\zeta)-\mathrm{i}\,b(\zeta), a(\zeta')+\mathrm{i}b(\zeta')) = \mathrm{Cov}_Q(a(\zeta),a(\zeta')) - \mathrm{iCov}_Q(b(\zeta),a(\zeta')) + \mathrm{iCov}_Q(a(\zeta),b(\zeta')) + \mathrm{Cov}_Q(b(\zeta),b(\zeta'))$. But since a and b are independent, we get $\mathrm{Cov}_Q(b(\zeta),a(\zeta')) = \mathrm{Cov}_Q(a(\zeta),b(\zeta')) = 0$, and then $\mathrm{Cov}_Q(\hat{x}(-\zeta),\hat{x}(\zeta')) = \mathrm{Cov}_Q(a(\zeta),a(\zeta')) + \mathrm{Cov}_Q(b(\zeta),b(\zeta'))$. The same derivations for the rest three terms, and we find

$$\begin{aligned}
&\mathrm{Cov}_Q(x(t),x(t')) \\
&= \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(\mathrm{i}\,\zeta t)\exp(\mathrm{i}\,\zeta' t')\,[\mathrm{Cov}_Q(a(\zeta),a(\zeta')) - \mathrm{Cov}_Q(b(\zeta),b(\zeta'))] \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(-\mathrm{i}\,\zeta t)\exp(\mathrm{i}\,\zeta' t')\,[\mathrm{Cov}_Q(a(\zeta),a(\zeta')) + \mathrm{Cov}_Q(b(\zeta),b(\zeta'))] \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(\mathrm{i}\,\zeta t)\exp(-\mathrm{i}\,\zeta' t')\,[\mathrm{Cov}_Q(a(\zeta),a(\zeta')) + \mathrm{Cov}_Q(b(\zeta),b(\zeta'))] \\
&+ \int_0^{+\infty} \mathrm{d}\zeta \int_0^{+\infty} \mathrm{d}\zeta' \exp(-\mathrm{i}\,\zeta t)\exp(-\mathrm{i}\,\zeta' t')[\mathrm{Cov}_Q(a(\zeta),a(\zeta')) - \mathrm{Cov}_Q(b(\zeta),b(\zeta'))].
\end{aligned}$$

The covariances can be read from $p(x)$. To do so, we discretize the expression in the exponential of $p(x)$, as

$$-\beta \sum_{n=0}^{+\infty} \Delta\zeta\,(\zeta_n^2+\omega^2)\,[a^2(\zeta_n)+b^2(\zeta_n)],$$

where $\zeta_n := n\Delta\zeta$. As a normal distribution, the covariance $\mathrm{Cov}_Q(a(\zeta_n),a(\zeta_{n'})) = \delta_{n,n'}/[2\beta\Delta\zeta\,(\zeta_n^2+\omega^2)]$. When $\Delta\zeta\to 0$, $\delta_{n,n'}/\Delta\zeta$ becomes Dirac's delta function $\delta(\zeta-\zeta')$. This can be realized as follow. For any function f, we have $f(\zeta_{n'}) = \sum_n f(\zeta_n)\,\delta_{n,n'} = \sum_n \Delta\zeta f(\zeta_n)\,(\delta_{n,n'}/\Delta\zeta)$, which becomes $\int \mathrm{d}\zeta f(\zeta)\,\delta(\zeta-\zeta')$ as $\Delta\zeta\to 0$. Thus, $(\delta_{n,n'}/\Delta\zeta)$ is recognized as $\delta(\zeta-\zeta')$. So, we get $\mathrm{Cov}_Q(a(\zeta),a(\zeta')) = \delta(\zeta-\zeta')/[2\beta\,(\zeta^2+\omega^2)]$. The same, $\mathrm{Cov}_Q(b(\zeta),b(\zeta')) = \delta(\zeta-\zeta')/[2\beta\,(\zeta^2+\omega^2)]$. Plugging these to $\mathrm{Cov}_Q(x(t),x(t'))$, the first and the last lines vanish, and we arrive at

$$\mathrm{Cov}_Q(x(t),x(t')) = \int_0^{+\infty} \mathrm{d}\zeta\,\frac{\exp(\mathrm{i}\,\zeta\,(t-t'))}{\beta\,(\zeta^2+\omega^2)} + \int_0^{+\infty} \mathrm{d}\zeta \frac{\exp(-\mathrm{i}\,\zeta\,(t-t'))}{\beta\,(\zeta^2+\omega^2)} = \int_{\mathbb{R}} \mathrm{d}\zeta\,\frac{\exp(\mathrm{i}\,\zeta\,(t-t'))}{\beta\,(\zeta^2+\omega^2)}.$$

This integral, which results in the Yukawa potential, is very tricky. The key is noticing the following integral

$$\int_0^x \mathrm{d}u \exp(-(\omega\pm\mathrm{i}\zeta)\,u) = \frac{1-\exp(-(\omega\pm\mathrm{i}\zeta)\,x)}{\omega\pm\mathrm{i}\zeta}.$$

When $\omega > 0$, let $x \to +\infty$, we find

$$\int_0^{+\infty} \mathrm{d}u \exp(-(\omega \pm \mathrm{i}\zeta)\, u) = \frac{1}{\omega \pm \mathrm{i}\zeta}\Big[1 - \lim_{u\to+\infty} \exp(-(\omega \pm \mathrm{i}\zeta)\, u)\Big] = \frac{1}{\omega \pm \mathrm{i}\zeta}.$$

This relation helps convert the denominator in the covariance to exponential, as

$$\frac{1}{\zeta^2+\omega^2} = \frac{1}{2\omega}\left[\frac{1}{\omega+\mathrm{i}\zeta} + \frac{1}{\omega-\mathrm{i}\zeta}\right] = \frac{1}{2\omega}\left[\int_0^{+\infty} \mathrm{d}u \exp(-(\omega+\mathrm{i}\zeta)\, u) + \int_0^{+\infty} \mathrm{d}u \exp(-(\omega-\mathrm{i}\zeta)\, u)\right].$$

Plugging into the covariance, and exchanging the integrals $\int_{\mathbb{R}} \mathrm{d}\zeta$ and $\int_0^{+\infty} \mathrm{d}u$, we get

$$\int_{\mathbb{R}} \mathrm{d}\zeta \frac{\exp(\mathrm{i}\,\zeta\,(t-t'))}{\zeta^2+\omega^2} = \frac{1}{2\omega}\int_0^{+\infty} \mathrm{d}u \exp(-\omega u)\left[\int_{\mathbb{R}} \mathrm{d}\zeta \exp(\mathrm{i}\,\zeta\,(t-t'-u)) + \int_{\mathbb{R}} \mathrm{d}\zeta \exp(\mathrm{i}\,\zeta\,(t-t'+u))\right].$$

Integration of ζ turns to be Dirac's delta functions, which result in $(1/2\omega)\exp(-\omega\,|t-t'|)$. So, $\mathrm{Cov}_Q(x(t), x(t')) = (2\beta\omega)^{-1}\exp(-\omega\,|t-t'|)$ with $\omega > 0$, from which the Pearson coefficient reads

$$\mathrm{Pearson}_Q(x(t), x(t')) = \exp(-\omega\,|t-t'|),$$

which decreases exponentially with $|t-t'|$. The decay rate is $1/\omega$.

Only in the limit $\omega \to 0$ is the system scale-free, which means the information can propagate to the edge of the system (which is infinity in our case). This is plausible, since the action becomes scale-invariant in and only in the same limit, where there is no dimensional constant (ω has dimension $[t]^{-1}$). But, we cannot say that a stochastic system is scale-free if and only if the action is scale-invariant. As an example, the Wilson-Fisher fixed point is an instance where the stochastic system is scale-free but the action is not scale-invariant (see this post). We will examine this in section TODO.

4.5 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 parametrized distribution with parameters θ. From its density function, $p(\cdot, \theta)$, we get a parameterized action $S(\cdot, \theta)$ such that

$$p(x, \theta) = \exp(-S(x, \theta))\,/\,Z(\theta), \tag{23}$$

where $Z(\theta) = \int_{\mathcal{X}} \mathrm{d}x \exp(-S(x, \theta))$ for ensuring the normalization $\int_{\mathcal{X}} \mathrm{d}x p(x, \theta) = 1$. This is consistent with the action defined by equation 22, except that the action here is parameterized, and that we omit the constant β since it is irrelevant throughout this section.

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 $H(Q, P(\theta)) := \int_{\mathcal{X}} \mathrm{d}x q(x) \ln(q(x)/p(x, \theta))$. 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 23 into $H(Q, P(\theta))$,

$$H(Q, P(\theta)) = \int_{\mathcal{X}} \mathrm{d}x\, q(x) \ln q(x) + \int_{\mathcal{X}} \mathrm{d}x\, q(x)\, S(x, \theta) + \int_{\mathcal{X}} \mathrm{d}x\, q(x) \ln Z(\theta).$$

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

$$L(\theta) := \int_{\mathcal{X}} \mathrm{d}x\, q(x)\, S(x, \theta) + \ln Z(\theta).$$

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

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

The first term is recognized as $-\mathbb{E}_Q[\partial S / \partial \theta^\alpha]$. For the second term, since $Z(\theta) = \int_{\mathcal{X}} \mathrm{d}x \exp(-S(x, \theta))$, we have

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

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

$$-\frac{\partial L}{\partial \theta^\alpha}(\theta) = \mathbb{E}_{P(\theta)}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right] - \mathbb{E}_Q\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right]. \tag{24}$$

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

$$\mathbb{E}_{P(\theta_\star)}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta_\star)\right] = \mathbb{E}_Q\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta_\star)\right]. \tag{25}$$

It can be read from equation 24 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 2 illustrates, this way of optimization will site real world data onto local minima of $S(\cdot, \theta)$, *in statistical sense*.

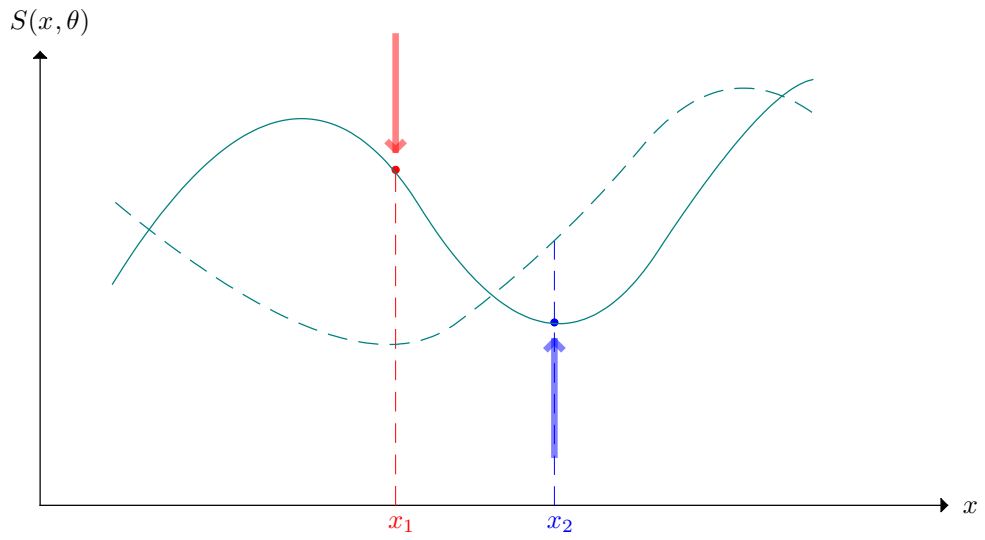

Figure 2. This figure illustrate how $\min_\theta L(\theta)$ will site a real world datum onto a local minimum of $S(\cdot, \theta)$. The green curve represents the current not-yet-optimized $S(\cdot, \theta)$. The x_1 (red point) is a real world datum while x_2 (blue point), which is currently a local 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.

In this way, we find an analytical distribution $P(\theta)$ that approximates the empirical distribution Q. The $S(\cdot, \theta)$ that defines $P(\theta)$ describes the interaction between the different components of an entity. This entity may be of physics, like a collection of particles. But it can also be words, genes, flock of birds, and so on.

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. The parameterized action S can be expressed by a neural network. Then, iterating by equation 24 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 24, we can employ Monte-Carlo simulation with the transition rate satisfying detailed balance condition with $P(\theta)$ as the stationary distribution. For continuous random variables, Monte-Carlo simulation with Langevin dynamics (section 3.7) is efficient; and for discrete random variables, Metropolis-Hastings (section 2.7) is available.

4.6 Structures in Nature May Arise from Least-Action Principle

There are many structures in nature. The structure of vascular system is a simple instance. A more complicated structure appears in the bases along chromosome. Why do these structures arise in nature?

Early in 1997, physicist Geoffrey West, ecologist James Brown, and biologist Brian Enquist proposed a theory (now it is called WBE theory) that explains how the fractal structures arise in vascular system of mammals.[14] To do so, they *derived* an objective that quantifies the cost of transporting blood. They found that the fractal structure of vascular appears naturally by minimizing this cost. Also arises the power-law relationship between the basal metabolic rate and the body size of mammal, which was first observed by Max Kleiber in 1930 and now named by Kleiber's law. Later, they applied their theory to many areas that have no superficial relationship with biology, such as gross domestic product of city. They successfully predicted some observed quantities in these areas.

Inspired by WBE theory, we regard the cost as an action. Instead of deriving a cost/action as WBE does, we can use the technique declared in section 4.5 to reveal one if we have obtained sufficiently many observed data. In machine learning perspective, data fitting is also seen as pattern mining. It reveals the statistically significant patterns hidden in the data. These patterns are the structures frequently appear in nature, and they locate in the minima of an objective, as WBE theory claimed, an action.

An interesting aspect of WBE theory is that the quantitative results obtained by minimizing the cost in one system are also held by a large variety of systems in nature. For example, different systems may share the same power-law index. This property is called **universality**. Where does univerality come from?

In 1975, physcist Mitchell Feigenbaum computed two constants, now named as Feigenbaum constants, when he was studying the logistic map. Then in the late of 1970s, physicsts found that Feigenbaum constants also appear in many other areas such as turbulence and Mandelbrot set: Feigenbaum constants are universal. Feigenbaum himself gave a "proof" of how this universality appears. The technique he used was invented by his collage in Cornell University, Kenneth Wilson, called renormalization group. With this technique, Feigenbaum constructed a functional iterative equation, and found his constants as the Taylor coefficients of the non-trivial fixed point of the functional iterative equation. But, Feigenbaum said little about where this functional iterative equation comes from. He neither gave a rigerous derivation of the equation, nor argued why this equation holds also for other systems.

Generally, universality comes from a "complex" system, a system whose configuration (defined in section 4.2) has a large number of components, such as starling flocks or ant colony. In such systems, each component can only interact with several "neighbors". But, when a local perturbation (for example, caused by a predator) appears, its information soon propagates throughout the whole system, and the system reacts to the perturbation as a large complex organism, which is where the name "complex" emerges. Phenomenon that information propagates throughout the whole system without decay is called **criticality**. This is important for starling flocks or ant colony to survive, and the cost will be strongly related to the appearance of criticality.

TODO

14. *A General Model for the Origin of Allometric Scaling Laws in Biology*. DOI: 10.1126/SCIENCE.276.5309.122

4.7 Example: Action of Feed-forward Neural Network and More

In deep learning, a feed-forward network is a supervised model that computes the output $y \in \mathbb{R}^{n_L}$ from input $x \in \mathbb{R}^{n_0}$. To do so, it computes a series of intermediate quantities called hidden variables $(h_1, \ldots, h_{L-1})$ with $h_l \in \mathbb{R}^{n_l}$ iteratively by

$$h_{l+1} = f(h_l, \theta_{l+1}), \tag{26}$$

where $f(\cdot, \theta_l)\colon \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$ with θ_l its parameters. By denoting $h_0 := x$ and $h_L := y$, we have $l \in \{0, \ldots, L\}$.

To investigate feed-forward neural network under probabilistic perspective, we have to construct the distribution of random variables $(H_0, \ldots, H_L)$ where H_l denotes the output of the l-th layer, H_0 the input, and H_L the output. We assume that H_{l+1} obeys the normal distribution with mean given by the deterministic iteration and a fixed variance ϵ (shared for all components of H_{l+1}), as

$$H_{l+1} \sim \mathcal{N}(f(H_l, \theta_{l+1}), \epsilon), \tag{27}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. It has the conditional density function

$$q(h_{l+1}|h_l) = \left(\frac{1}{\sqrt{(2\pi)\,\epsilon}}\right)^{n_{l+1}} \times \exp\left(-\frac{(h_{l+1} - f(h_l, \theta_{l+1}))^2}{2\epsilon}\right).$$

To obtain the whole density function, we first notice that $q(h_1|h_0)\, q(h_0) = q(h_0, h_1)$. And since $q(h_2|h_1)$ is not explicitly dependent on h_0, $q(h_2|h_1) = q(h_2|h_0, h_1)$ holds for any h_0. Then, $q(h_2|h_1) q(h_1|h_0)\, q(h_0) = q(h_2|h_0, h_1)\, q(h_0, h_1) = q(h_0, h_1, h_2)$. Repeating this step, we will find

$$q(h_0, \ldots, h_L) = q(h_L|h_{L-1}) \cdots q(h_1|h_0)\, q(h_0).$$

Plugging in $q(h_{l+1}|h_l)$ and explicitly put θ into $q(h_0, \ldots, h_L)$, we arrive at

$$-\ln q(h_0, \ldots, h_L, \theta) = \frac{1}{2\epsilon} \sum_{l=0}^{L-1} (h_{l+1} - f(h_l, \theta_{l+1}))^2 + \text{const.}$$

By equation 22 (setting $\beta = 1$), the action of feed-forward neural network is recognized as

$$S(h, \theta) = \frac{1}{2\epsilon} \sum_{l=0}^{L-1} (h_{l+1} - f(h_l, \theta_{l+1}))^2.$$

We are to compare this action with that appearing in classical mechanics. To do so, we introduce $g(x, \theta) := (f(x, \theta) - x)/\epsilon$, thus $h_{l+1} = h_l + \epsilon\, g(h_l, \theta_{l+1})$. It is recognized as residual structure in deep learning, which was proposed by Kaiming He and others in 2015 for dealing with the issues caused by increasing the number of layers, L. So, equivalently,

$$-\ln q(h_0, \ldots, h_L, \theta) = \frac{\epsilon}{2} \sum_{l=0}^{L-1} \left(\frac{h_{l+1} - h_l}{\epsilon} - g(h_l, \theta_{l+1})\right)^2.$$

Comparing with classical mechanics, we can interpret ϵ as a tiny time interval and $(h_{l+1} - h_l)/\epsilon$ as "velocity". It motives us to consider its continuous version

$$-\ln q(h) = \frac{1}{2} \int_0^1 \mathrm{d}t\, [\dot{h}(t) - g(h(t), t)]^2,$$

where we have replaced ϵ by $\mathrm{d}t$, $(h_0, \ldots, h_L)$ by $h(t)$ with $t \in [0, 1]$, and absorbed $\theta(t)$ into g so as to make it comparable with classical mechanics. To get the Lagrange L, which is defined by $S(h) = \int_0^1 \mathrm{d}t\, L(h(t), \dot{h}(t), t)$, we expand the integrand and find

$$L(h, \dot{h}, t) = \frac{1}{2}\dot{h}^2 - \dot{h}\, g(h, t) + \frac{1}{2} g^2(h, t).$$

Euler-Lagrange equation $(\mathrm{d}/\mathrm{d}t)\partial L/\partial \dot{h} = \partial L/\partial h$ gives

$$\ddot{h} = g(h, t) + \partial_t g(h, t).$$

When the feed-forward neural network shares weights crossing layers, $g(h, t)$ will be independent of t, and we will get $\ddot{h} = g(h)$, a typical second-order equation of motion with the "force" g.

This action, however, can be applied to many other areas that are irrelevant to machine learning, since equation 26 is much more general. The assumption of Gaussianity 27 together with interpreting its variance as time interval reminds us of the Langevin dynamics (see section 3.3) where the Wiener process is Gaussian with variance $\mathrm{d}t$. TODO