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 $p(y) := \int_{\mathcal{Z}} \mathrm{d}z p(y,z)$ and $p(z) := \int_{\mathcal{Y}} \mathrm{d}y p(y, z)$ if X is of continuum, and the same for mass function. Notice that we have omitted the subscript Y in p_Y (and the same for p_Z) since the y in $p(y)$ has clearly indicated this. We **marginalize** Z so as to get P_Y.

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

1.2 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

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

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.

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.

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 $P_Y, \ldots, Q_Z$ 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.

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,

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

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.

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

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

where $\delta(x-y)$ indicates Kronecker's delta function when $\mathcal{X}$ is discrete, or Dirac's delta function when $\mathcal{X}$ is continuous. Delta function has the property that $\int_{\mathcal{X}} \mathrm{d}x\, \delta(x-y)\, f(x) = f(y)$ for any f.

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

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

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

$$\frac{\partial p}{\partial t}(x,t) = \int_{\mathcal{X}} \mathrm{d}y\, r(x,y)p(y,t). \tag{4}$$

Since we only deal with Markovian stochastic process throughout this note, when referring to master equation, we mean equation 4. And to discrete time master equation, equation 5:

$$p(x,t+\Delta t) = \int_{\mathcal{X}} \mathrm{d}y\, q_{\Delta t}(x,y)p(y,t). \tag{5}$$

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 4 to calculate $\partial^n p/\partial t^n$ for any n. For $n = 2$, by inserting master equation 4 (to the blue term), we have

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

We then insert master equation 4 again (to the green term), and find

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

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.

On the other hand, Taylor expand the both sides of equation 5 by Δt gives, at $(\Delta t)^{n+1}$ order,

$$\frac{\partial^{n+1} p}{\partial t^{n+1}}(z,t) = \int_{\mathcal{X}} \mathrm{d}x\, q_0^{(n+1)}(z|x) p(x,t),$$

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

$$q_{\Delta t}^{(n)}(x|y) := \lim_{s \to \Delta t} \frac{\mathrm{d}^n q_s}{\mathrm{d}s^n}(x|y).$$

So, by equaling the two expressions of $(\partial^{n+1} p/\partial t^{n+1})(z,t)$, we find

$$\int_{\mathcal{X}} \mathrm{d}x \left[q_0^{(n+1)}(z|x) - \int_{\mathcal{X}} \mathrm{d}y_1 \cdots \int_{\mathcal{X}} \mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1}) \cdots r(y_1,x) \right] p(x,t) = 0$$

For $n = 1, 2, \ldots$. This holds for all $p(x,t)$, thus

$$q_0^{(n+1)}(z|x) = \int_{\mathcal{X}} \mathrm{d}y_1 \cdots \int_{\mathcal{X}} \mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1}) \cdots r(y_1,x).$$

Recalling that $q_{\Delta t}(z|x) = \delta(z-x) + r(z,x)\,\Delta t + o(\Delta t)$, we have the Taylor expansion of $q_{\Delta t}$, as[5]

$$\begin{aligned} q_{\Delta t}(z|x) = &\ \delta(z-x) \\ + &\ (\Delta t)\, r(z,x) \\ + &\ \frac{(\Delta t)^2}{2!}\int_{\mathcal{X}} \mathrm{d}y r(z,y)\, r(y,x) \\ + &\ \cdots \\ + &\ \frac{(\Delta t)^{n+1}}{(n+1)!}\int_{\mathcal{X}} \mathrm{d}y_1 \cdots \int_{\mathcal{X}} \mathrm{d}y_n\, r(z,y_n)\, r(y_n,y_{n-1}) \cdots r(y_1,x) \\ + &\ \cdots. \end{aligned} \tag{6}$$

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

This may be a little weird at the first sight. For example, consider $q'_{\Delta t}(y|x) := q_{\Delta t}(y|x) + f(y, x)\,\Delta t^2$, 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 the (continuous time) master equation 4 with the same r as that of $q_{\Delta t}$. So, we should have $q'_{\Delta t} = q_{\Delta t}$, which means f is not free, but should vanish.

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

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

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

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

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

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

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

$$q_{\Delta t+\Delta t'}(z|x) = \int_{\mathcal{X}} \mathrm{d}y q_{\Delta t'}(z|y)\, q_{\Delta t}(y|x). \tag{7}$$

This provides an addition restriction to the transition density.

---

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.

Interestingly, obeying equation 7 is sufficient for $q_{\Delta t}$ to satisfy equation 6. Precisely, let $q_{\Delta t}(x|y)$ is a function that is smooth on Δt with $q_0(x|y) = \delta(x-y)$, we are to show that, if $q_{\Delta t}$ satisfies equation 7, then it will obey equation 6. To do so, we take derivative on equation 7 by $\Delta t'$ at $\Delta t' = 0$, resulting in

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

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

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

and on the left hand side,

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

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

$$q^{(1)}_0(y|x) = r(y,x),$$

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

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

Iteratively at $(\Delta t)^{n+1}$ order, we will find

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

again. And this implies equation 6. So, we conclude this paragraph as follow: *obeying equation 7 is the sufficient and essential condition for a function $q_{\Delta t}(x|y)$, which is smooth on Δt with $q_0(x|y) = \delta(x-y)$, to satisfy equation 6; additionally, if $q_{\Delta t}(x|y)$ is non-negative and normalized on x (namely, $\int_{\mathcal{X}} \mathrm{d}x\, q_{\Delta t}(x|y) = 1$), then $q_{\Delta t}$ is a transition density.*

2.4 Detailed Balance Provides Stationary Distribution

Let Π a stationary solution of master equation 4. 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

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

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

$$r(x,y)\, \pi(y) = r(y,x)\pi(x), \tag{8}$$

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

Interestingly, for a transition rate r that satisfies detailed balance condition 8, the transition density $q_{\Delta t}$ generated by r using equation 6 satisfies a similar relation

$$q_{\Delta t}(x|y)\, \pi(y) = q_{\Delta t}(y|x)\pi(x). \tag{9}$$

To see this, consider the third line in equation (6), 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 6 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

$$H(P(t),\Pi) = \int_{\mathcal{X}} \mathrm{d}x\, p(x,t) \ln \frac{p(x,t)}{\pi(x)}. \tag{10}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t} H(P(t),\Pi) = -\frac{1}{2} \int_{\mathcal{X}} \mathrm{d}x \int_{\mathcal{X}} \mathrm{d}y\, r(x,y)\,\pi(x) \left(\frac{p(x,t)}{\pi(x)} - \frac{p(y,t)}{\pi(y)} \right) \left(\ln \frac{p(x,t)}{\pi(x)} - \ln \frac{p(y,t)}{\pi(y)} \right). \tag{11}$$

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 4, 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 8 holds for Π and r, then for any time-dependent distribution $P(t)$ with the same alphabet (at one time) evolved by the master equation 4, $P(t)$ will monotonically and constantly relax to Π.*

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

2.6 Monte-Carlo Simulation and Guarantee of Relaxation

How to numerically simulate the evolution of master equation 4 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 8. (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 5 instead, and detailed balance condition 9 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 4, $(\partial/\partial t)\ln p(x,t) = \int_{\mathcal{X}} \mathrm{d}x\, r(x,y)\, p(y,t)/p(x,t)$. For ΔH_2, we insert equation 11 after Taylor expanding $q_{\Delta t}$ by Δt, and obtain

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

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

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

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

This bound is proportional to the difference between $P(t)$ and Π (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

$$r(x,y)=\min\left(1,\frac{\gamma(y,x)\,\pi(x)}{\gamma(x,y)\,\pi(y)}\right)\gamma(x,y). \tag{12}$$

This transition rate connects every two elements in $\mathcal{X}$ (since $\gamma(y,x)>0$ for each $x\neq y$). In addition, together with π, it satisfies the detailed balance condition 8. 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 4, any initial distribution will finally relax to the stationary distribution* Π.

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

$$q(x|y):=\min\left(1,\frac{g(y|x)\,\pi(x)}{g(x|y)\,\pi(y)}\right)g(x|y), \tag{13}$$

where g is positive-definite on $\mathcal{X}$. Notice that, in equation 13 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 13 to numerically simulate master equation 4.* 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 8 holds. Actually, equation 8 *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,

$$\ln\pi(y)=\ln\pi(x)+\int_P \mathrm{d}s\,\rho(s), \tag{14}$$

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

$$\oint_C \mathrm{d}s\,\rho(s)=0. \tag{15}$$

It means every path along two connected elements in $\mathcal{X}$ is equivalent. If the condition 15 holds, we can simplify the notation in equation 14 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 15 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 15, $\rho(y,x)=\int_x^{x_0}\mathrm{d}s\,\rho(s)+\int_{x_0}^{y}\mathrm{d}s\,\rho(s)$ (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 8. 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 15 is *essential and sufficient for the existence of π that holds the detailed balance condition 8*. 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 15 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 convention**. That is, we omit the sum notation for the duplicated indices as long as they are "balanced". For example, $x_\alpha y^\alpha$ represents $\sum_\alpha x_\alpha y^\alpha$. The α appears twice in the expression, once in subscript (the x_α) and once in superscript (the y^α), for which we say indices are balanced. Expression like $x_\alpha y_\alpha$, however, does not represent a summation over α, because indices are not balanced (both are subscript). A more complicated example is $\partial_\alpha A^\alpha_\beta x^\beta$, which means $\sum_\alpha \sum_\beta \partial_\alpha A^\alpha_\beta x^\beta$.

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 4 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")

$$K^{\alpha^1\cdots\alpha^k}(x) := \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_k})\, \omega(x,\epsilon). \tag{16}$$

Finally, we arrive at

$$\frac{\partial p}{\partial t}(x,t) = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \left(\frac{\partial}{\partial x^{\alpha_1}}\cdots\frac{\partial}{\partial x^{\alpha_k}}\right) [K^{\alpha^1\cdots\alpha^k}(x)\, p(x,t)]. \tag{17}$$

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

3.3 From Brownian Motion to Central Limit Theorem

One important application of Kramers–Moyal expansion is Brownian motion. In 1827, botanist Robert Brown noticed that pollen particles automatically shakes in water. This phenomenon was first explained by Albert Einstein in 1905. He argued that the pollen particles are constantly stricken by water molecules.

We are to quantitatively determine how a pollen particle moves in water. The random movement of a pollen particle can be characterized by a transition density $q_{\Delta t}(x+\epsilon|x)$, where the pollen particle transits from x to $x+\epsilon$ during time interval Δt. For this transition density, we make two assumptions. The first comes from the observation that the pool under the microscope of Brown is much broader than the diameter of water molecule, and the temperature of water is uniform, so that a water molecule cannot distinguish where it locates, just like a boat floating on the ocean, because every place is the same. It indicates that the transition is homogeneous, namely, $q_{\Delta t}(x+\epsilon|x)$ does not depend on x. It, then, implies that the transition rate $r(x+\epsilon,x)$ is independent of x. This landscape also gives the other assumption that every direction is the same too: the transition is also isotropic. It indicates that $\int_{\mathbb{R}^n} \mathrm{d}x\, q_{\Delta t}(x+\epsilon|x)\, \epsilon^\alpha = 0$ for each α, since the water molecule cannot distinguish the direction $-\epsilon^\alpha$ from ϵ^α. With these two assumptions, the Kramers–Moyal expansion 17 becomes

$$\frac{\partial p}{\partial t}(x,t) = \sum_{k=2}^{+\infty} \frac{(-1)^k}{k!} K^{\alpha^1\cdots\alpha^k} \left(\frac{\partial}{\partial x^{\alpha_1}}\cdots\frac{\partial}{\partial x^{\alpha_k}}\right) p(x,t),$$

where the k starts at 2 (since the assumption $\int_{\mathbb{R}^n} \mathrm{d}x\, q_{\Delta t}(x+\epsilon|x)\, \epsilon^\alpha = 0$ implies $K^\alpha(x) = 0$) and the $K^{\alpha^1\cdots\alpha^k}(x) := \int_{\mathbb{R}^n} \mathrm{d}\epsilon\, (\epsilon^{\alpha_1}\cdots\epsilon^{\alpha_k})\, r(x+\epsilon|x)$ is constant now (because the $r(x+\epsilon,x)$ is independent of x).

Now, we are to examine the K carefully. It is determined by transition rate, that is, by the transition where Δt is infinitesimal (at least sufficiently small). In this situation, there will be at most one water molecule that strikes the pollen particle, so that the typical scale of ϵ is extremely tiny (much smaller than the capacity of Brown's microscope). So, we have $K^{\alpha_1\cdots\alpha_k} \gg K^{\alpha_1\cdots\alpha_k\alpha_{k+1}}$ for any $k \geqslant 2$ since the later contains more ϵ ($k=1$ is not so because $K^\alpha = 0$). This leads to a valid approximation

$$\frac{\partial p}{\partial t}(x,t) = \frac{1}{2} K^{\alpha\beta} \left(\frac{\partial}{\partial x^\alpha}\frac{\partial}{\partial x^\beta}\right) p(x,t),$$

where only the leading term $k=2$ remains. This equation is the famous heat equation, first developed by French mathematician Joseph Fourier in 1822. For initial value $p(x,0)$, it has the solution

$$p(x,t)=\frac{1}{\sqrt{(2\pi t)^n \det(K)}}\int_{\mathbb{R}^n}\mathrm{d}y\exp\left(-\frac{1}{2t}(K^{-1})_{\alpha\beta}\,(x^\alpha-y^\alpha)\,(x^\beta-y^\beta)\right)p(y,0),$$

where the factor $1/\sqrt{\cdots}$ comes from normalization $\int_{\mathbb{R}^n}\mathrm{d}x\,p(x,t)=1$. Recall the (discrete time) master equation 5, $p(x,\Delta t)=\int_{\mathbb{R}^n}\mathrm{d}y\,q_{\Delta t}(x|y)\,p(y,0)$. The transition rate of pollen particle can be readily read out as

$$q_{\Delta t}(x+\epsilon|x)=\frac{1}{\sqrt{(2\pi\Delta t)^n \det(K)}}\exp\left(-\frac{1}{2\Delta t}(K^{-1})_{\alpha\beta}\,\epsilon^\alpha\,\epsilon^\beta\right). \tag{18}$$

The phenomenon that this transition density describes is called **Brownian motion**. Even though the techniques used for deriving this transition density had been mature when Brown first observed this phenomenon, but almost one hundred years after Brown's discover, in 1918, Norbert Wiener first constructed a complete mathematical theory for this stochastic process. So, it is also called **Wiener process**.

The transition rate $q_{\Delta t}(x+\epsilon|x)$ can be seen as an accumulation of a series tiny transitions, each is caused by one strike from a water molecule. The strike obeys a distribution which is identical (each water molecule behaves in the same way, as a result of homogeneity) and independent (since each strike is individual) with zero mean (as a result of isotropy). This distribution, however, is unknown. Although, we find that the accumulative effect always obeys a normal distribution. We can abstract this and conclude a corollary as follow.

Corollary 2. *For any independently identically distributed n-dimensional random variables (X_1,..., X_N) with zero mean (thus each X_i is one strike), the accumulation $Y:=X_1+\cdots+X_N$ tends to obey a normal distribution as N is large enough.*

Each X_i can be seen as a strike by water molecule. Further, the mean of Y can be calculated by the linearity of expectation, as $\mathbb{E}[Y]=\mathbb{E}[X_1]+\cdots+\mathbb{E}[X_N]=0$. And because of independency, we have $\mathbb{E}[Y^\alpha Y^\beta]=\mathbb{E}[X_1^\alpha X_1^\beta]+\cdots+\mathbb{E}[X_N^\alpha X_N^\beta]$. Let $\Sigma^{\alpha\beta}:=\mathbb{E}[X_i^\alpha X_i^\beta]$, which is the same for all i because X_is are identical, we find $\mathbb{E}[Y^\alpha Y^\beta]=N\Sigma^{\alpha\beta}$. This is the **central limit theorem**, the most famous theorem in probability theory. Now, we have found for central limit theorem a physical description, the Brownian motion, and found it as a corollary of Kramers–Moyal expansion.

3.4 Langevin Process Arises in the Difference of Scales

There are many levels of scale in Nature. From the lifetime of universe to the lifetime of human. From the movement of a bird to the movement of molecule. We are to formulate the general mathematical description for the system in which multiple scales are involved.

A typical example is the Brownian motion described in section 3.3. Of course, each water molecule moves the pollen particle in such a tiny distance that cannot be observed by a microscope made in the 19th century. What Brown noticed was not a pollen particle shaken by a single water molecule, but an accumulation of strikes by a large group of water molecules. So, this phenomenon involves two different scales: the scale of pollen particles, and the scale of movement of water molecules, which is much smaller than the frontier.

If we replace the pollen particles by paramecia, then the scales remain. In the perspective of water molecule, homogeneity and isotropy still hold. So, the contribution from the constant striking of water molecules obeys the transition density of Wiener process 18. But in the perspective of paramecium, both homogeneity and isotropy break. Unlike pollen particle, a paramecium can swim along a direction (maybe, there is food on this direction), thus isotropy breaks. In the perspective of the paramecium, which is much larger than a water molecule, the pool is not like an ocean anymore, but a pond. So, after arriving at another place in the pool, it can feel the change of environment (such as the temperature of water), thus homogeneity breaks.

This pattern arises in many areas of Nature in which two scales coexist simultaneously: one scale is smaller, being homogeneous and isotropic, while the other is greater, breaking homogeneity and isotropy. The greater scale obeys a deterministic behavior, characterized by a dynamical system $\mathrm{d}x^\alpha/\mathrm{d}t = f^\alpha(x)$, or difference equation $x^\alpha_{i+1} = x^\alpha_i + f^\alpha(x_i)\,\Delta t$ for some small but still finite time interval Δt (the subscripts denote the iterative steps). While the smaller scale contributes to the movement of the greater one by the accumulative random effect ΔW^α_i, which is proven to obey a normal distribution with zero mean and covariance $K^{\alpha\beta}(x_i)\,\Delta t$ (proved in section 3.3). Notice that, since the homogeneity has broken at the greater scale, $K^{\alpha\beta}$ will explicitly depends on position x. So, the total effect is

$$x^\alpha_{i+1} = x^\alpha_i + f^\alpha(x_i)\,\Delta t + \Delta w^\alpha_i, \tag{19}$$

where Δw_i is sampled from the distribution of ΔW_i.

We are to determine the conditional probability of x_{i+1} given x_i, where the randomness of x_{i+1} comes from that of ΔW_i. We know that a linear combination of random variables that obey normal distribution also obeys a normal distribution. Then, since x_{i+1} is linear with Δw_i, we have X_{i+1} (the random version of x_{i+1}) will also obey a normal distribution when x_i is fixed, with the conditional density function on $\mathbb{R}^n$

$$q_{\Delta t}(x_{i+1}|x_i) := \frac{1}{\sqrt{(2\pi\Delta t)^n \det K(x_i)}}\exp\Big(-\frac{1}{2\Delta t}\,[K^{-1}(x_i)]_{\alpha\beta}\,[x^\alpha_{i+1} - x^\alpha_i - f^\alpha(x_i)\,\Delta t]\,[x^\beta_{i+1} - x^\beta_i - f^\beta(x_i)\,\Delta t]\Big). \tag{20}$$

When Δt is sufficiently small, $q_{\Delta t}$ can be approximately regarded as a transition density (the essential and sufficient condition for $q_{\Delta t}$ to be a transition density was discussed in section 2.3).[10] The corresponding Markov process is called **Langevin dynamics** or **Langevin process**. In many textures, it is written in

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

which is a formal re-formulation of equation 19.

3.5 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 $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ in **Schwarts space**, which is a functional space in which function 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 Schwarts space $S(\mathbb{R},\mathbb{R})$ (the first $\mathbb{R}$ represents for domain and the second for codomain). Then, we Taylor expand f by its variable

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

10. Is $q_{\Delta t}$ a transition density? In section ?, we have shown that $q_{\Delta t}$ is a transition density if and only if $q_{\Delta t+\Delta t'}(x|z) = \int_{\mathbb{R}^n}\mathrm{d}y\, q_{\Delta t'}(x|y)\,q_{\Delta t}(y|z)$. By inserting the $q_{\Delta t}$ of Langevin process, we find the integrand in the right hand side proportional to

$$q_{\Delta t'}(x|y) = \frac{1}{\sqrt{(2\pi\Delta t')^n \det K(y)}}\exp\Big(-\frac{1}{2\Delta t'}\,K^{-1}(y)\,(x-y-f(y)\,\Delta t')\,(x-y-f(y)\,\Delta t')\Big),$$

in which y appears in many places, including $\det K(y)$, $K^{-1}(y)$, and $f(y)$. Thus, that in the exponential is not quadratic on y. It is hard to expect that integrating over y will give a result that is proportional to

$$\exp\Big(-\frac{1}{2(\Delta t+\Delta t')}\,K^{-1}(z)\,[x-z-f(z)\,(\Delta t+\Delta t')]\,[x-z-f(z)\,(\Delta t+\Delta t')]\Big).$$

So, an educated guess is that $q_{\Delta t}$ is not a transition density, but just an approximation of some transition density when Δt is sufficiently small. Remark that, when $f=0$ and K is constant, it is straight-forward to show that $q_{\Delta t}$ is indeed a transition density.

These Gaussian integrals result in

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

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

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

Altogether,

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

as $\Delta t \to 0$ (for example, $\int_{\mathbb{R}^n} \mathrm{d}\epsilon\, q_{\Delta t}(x+\epsilon|x) [\, \epsilon^\alpha\, \epsilon^\beta\, \epsilon^\gamma\, \epsilon^\delta\, \partial_\alpha \partial_\beta \partial_\gamma \partial_\delta \varphi(0)] = \mathcal{O}(\Delta t^2) = o(\Delta t)$). 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

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

From the terms proportional to Δt, we recognize

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

Noticing the integration by parts[11]

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

and

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

we get

$$r(x+\epsilon, x) = -f^\alpha(x)\, \partial_\alpha \delta(\epsilon) + \frac{1}{2} K^{\alpha\beta}(x)\, \partial_\alpha \partial_\beta \delta(\epsilon). \tag{21}$$

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 $\varphi \in S(\mathbb{R}^n, \mathbb{R}^n)$, and calculate $\int_{\mathbb{R}^n} \mathrm{d}x\, \partial_\alpha \delta(-x)\, \varphi^\alpha(x)$. 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

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

---

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

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

Then, integration by parts gives $-\int_{\mathbb{R}^n}\mathrm{d}x\,\partial_\alpha[\delta(-x)]\,\varphi^\alpha(x)=\int_{\mathbb{R}^n}\mathrm{d}x\,\delta(-x)\,\partial_\alpha\varphi^\alpha(x)$. After inserting the relation $\delta(x)=\delta(-x)$, we arrive at $\int_{\mathbb{R}^n}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi^\alpha(x)=\partial_\alpha\varphi^\alpha(0)$. On the other hand, we have, by integration by parts, $-\int_{\mathbb{R}^n}\mathrm{d}x\,\partial_\alpha\delta(x)\,\varphi^\alpha(x)=\int_{\mathbb{R}^n}\mathrm{d}x\,\delta(x)\,\partial_\alpha\varphi^\alpha(x)=\partial_\alpha\varphi^\alpha(0)$. Altogether, we find $\int_{\mathbb{R}^n}\mathrm{d}x\,\partial_\alpha\delta(-x)\,\varphi^\alpha(x)=-\int_{\mathbb{R}^n}\mathrm{d}x\,\partial_\alpha\delta(x)\,\varphi^\alpha(x)$, for any $\varphi\in S(\mathbb{R}^n,\mathbb{R}^n)$. 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 3.8.

3.6 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 $K^\alpha(x)=f^\alpha(x)$, and those with order (the number of superscripts) higher than $K^{\alpha\beta}(x)$ are all vanishing (K is defined in section 3.2). For example, the integral $\int_{\mathbb{R}^n}\mathrm{d}\epsilon\,(\epsilon^\alpha\,\epsilon^\beta\,\epsilon^\gamma)\,q_{\Delta t}(x+\epsilon|x)=\mathcal{O}(\Delta t^{3/2})$, which can be easily realized by the estimation $\epsilon=\mathcal{O}(\sqrt{\Delta t})$. 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 17 reads

$$\frac{\partial p}{\partial t}(x,t)=-\partial_\alpha(f^\alpha(x)\,p(x,t))+\frac{1}{2}\partial_\alpha\partial_\beta(K^{\alpha\beta}(x)\,p(x,t)). \tag{22}$$

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.

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

The master equation of Langevin process (equation 22) has stationary solution Π which satisfies (since there is only one variable x, we use ∂ instead of ∇)

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

which means

$$f^\alpha(x)\,\pi(x)=\frac{1}{2}\partial_\beta(K^{\alpha\beta}(x)\,\pi(x))+\nu^\alpha(x), \tag{23}$$

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**.

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 $\partial_\alpha\partial_\beta\delta(-x)=\partial_\alpha\partial_\beta\delta(x)$.

3.8 Detailed Balance Condition of Langevin Process Lacks Source-Free Degree of Freedom

After discussing stationary distribution of Fokker-Planck equation (as a master equation), we continue investigate when will Langevin process relax an initial distribution to the stationary. By theorem 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 21, we get, for the left hand side,

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

and, for the right hand side,

$$\begin{aligned}
&r(x,x+\epsilon)\,\pi(x+\epsilon)\\
=&\,r((x+\epsilon)-\epsilon,x+\epsilon)\,\pi(x+\epsilon)\\
=&-f^\alpha(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\delta(-\epsilon)+\frac{1}{2}K^{\alpha\beta}(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\partial_\beta\delta(-\epsilon)\\
=&\,f^\alpha(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\delta(\epsilon)+\frac{1}{2}K^{\alpha\beta}(x+\epsilon)\,\pi(x+\epsilon)\,\partial_\alpha\partial_\beta\delta(\epsilon),
\end{aligned}$$

where in the last line, we have used $\partial_\alpha\delta(-x)=-\partial_\alpha\delta(x)$ and $\partial_\alpha\partial_\beta\delta(-x)=\partial_\alpha\partial_\beta\delta(x)$ (derived in the end of section 3.5).

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

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

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

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

The right hand side is a little complicated,

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

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

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

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

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

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

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

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

$$f^\alpha(x)\,\pi(x)=\frac{1}{2}\,\partial_\beta(K^{\alpha\beta}(x)\,\pi(x)).\tag{24}$$

Comparing with the stationary solution of Langevin process (equation 23), 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.