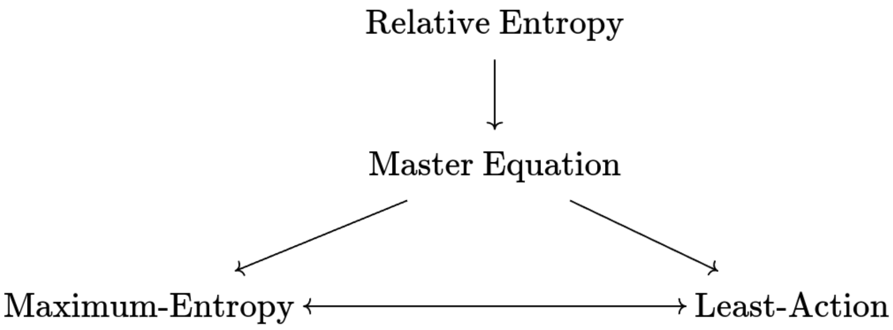

Figure 1. Scheme of sections. The solid arrow represents relation.

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 distributions**, 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 Variable

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

[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 $(p_1, \ldots, p_n)$;
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 not valid. 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. $H(P, Q) > 0$ if $P \neq Q$ 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.

[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.

The first condition employs the locality of H, which is thought as natural since H has be 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 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 $\partial_1 h(1,1)\int\mathrm{d}x p(x)\ln(p(x)/p(x))+C=0$. By Jensen's inequality, the integral $\int\mathrm{d}x p(x)\ln(p(x)/q(x))$ is non-negative, thus $\partial_1 h(1,1)>0$. The second equation results in $C=0$. 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.

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

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

As portion, $q_{t \rightarrow t'}$ cannot be negative, thus $q_{t \rightarrow t'}(x|y) \geqslant 0$ for each x and y in $\mathcal{X}$. We call $q_{t \rightarrow t'}$ 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 $\int_{\mathcal{X}} \mathrm{d}y q_{t \rightarrow t'}(y|x)\, p(x, t)$, and that transited to x, which is $\int_{\mathcal{X}} \mathrm{d}y q_{t \rightarrow t'}(x|y) p(y, t)$. Thus, we have

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

By inserting equation (1), we find

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

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

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

In addition, if the movement of sand is smooth, that is, a sand cannot disappear at one position and then suddenly reappear at another position far distant away, but has to move from one position to another by continuous movement, then $q_{t \rightarrow t'}$ 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 $r(x,y) > 0$ 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,

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

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

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

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

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

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

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

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. With an integral by part, they arrive at a negative definite expression, which means the monotonic reduction. This proof needs smooth structure on X, which is essential for integral by part. In this section, we provides a more generic alternative to the proof, for which smooth structure on X is unnecessary. It employs detailed condition instead of Fokker-Planck equation, which is a specific case of detailed balance (section 3.2).

2.6. Simulation of Master Equation and Guarantee of Relaxation

How to solve master equation 3? When the alphabet $\mathcal{X}$ is discrete and finite, where the desk becomes a finite two-dimensional array of lattices (recall section 2.2), we replace position by its index in the alphabet, like $x \rightarrow i$. Then, master equation becomes autonomous linear dynamical system: $(\mathrm{d} / \mathrm{d}t) p_i(t) = \sum_{j \in \mathcal{X}} r_{ij} p_j(t)$. It can be solved analytically, by investigating the eigen-system of matrix r.

[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 $z = y_n$.

When alphabet $\mathcal{X}$ is neither discrete nor finite, we can approximate the solution by using some properties of r. To do so, we expand $p(x,t)$ by a set of bases of Hilbert space, $\{\alpha_1,\alpha_2,\ldots\}$, as $p(x,t)=\sum_i \xi_i(t)\,\alpha_i(x)$, where ξ is the coefficient. The master equation turns to be

$$\sum_k \frac{\mathrm{d}\xi_k}{\mathrm{d}t}(t)\,\alpha_k(x)=\int_{\mathcal{X}}\mathrm{d}y r(x,y)\sum_j \xi_j(t)\,\alpha_j(x).$$

By inner product with $\alpha_i(x)$ on both side, we find

$$\frac{\mathrm{d}\xi_i}{\mathrm{d}t}(t)\,\alpha_i(x)=\sum_j \left[\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y r(x,y)\,\alpha_i(x)\,\alpha_j(y)\right]\xi_j(t).$$

If the r is specific (such as slowly varying) and the bases are properly chosen such that $\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y r(x,y)\,\alpha_i(x)\,\alpha_j(y)$ is negligible for any i and j greater than a number, the problem reduces to the case when $\mathcal{X}$ is discrete and finite.

In the worst situation, we have to solve master equation by numerical simulation. We simulate each sand, but replace its free will by a transition probability determined by transition rate r. Explicitly, we initialize the sand 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. We have to ensure that computer has a sampler that makes random sampling for $q_{\Delta t}(y|x)$ (as an example, see section 2.7). If r (or $q_{\Delta t}$), together with some stationary distribution Π, satisfies the detailed balance condition, then we *expect* that the simulation will iteratively decrease the difference between the distribution of the sands and the Π. We stop the iteration when they have been close enough.

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 evaluate 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 $\Delta t\to 0$.

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[8]

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

where, by master equation 3, $(\partial / \partial t) \ln p(x,t) = \int_{\mathcal{X}} \mathrm{d}x \, r(x,y) \, p(y,t) / p(x,t)$. The same for ΔH_1, which results in

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

where we have inserted equation 9. 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

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

This bound is proportional to the difference between $P(t)$ and Π (represented by the first factor). When $P(t)$ has approched 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 large 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.)

[8] The first line

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

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 I 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 I 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

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

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 $P := (p_0, \ldots, p_n)$ from x to y with $p_0 = x$ and $p_n = y$ (path and connectivity are defined in section 2.5), and define

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

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 Dynamics (TODO)

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 functions $p(x, t)$ and $q_{\Delta t}(x|y)$. This means, the conclusions in section 2 hold in this section, but the inverse is not guaranteed.

3.1. 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 $p(x, t)$ and $q_{\Delta t}(x|y)$ are smooth on x and y.

Now, the discrete time master equation 4 becomes

$$p(x, t + \Delta t) - p(x, t) = \int_{\mathbb{R}^n} \mathrm{d}y \, [q_{\Delta t}(x|y) \, p(y, t) - q_{\Delta t}(y|x) p(x, t)].$$

Let $\epsilon := x - y$ and $\omega(x, \epsilon) := q_{\Delta t}(x + \epsilon|x)$. We then have, for the first term,

$$\begin{aligned} & \int_{\mathbb{R}^n} \mathrm{d}y q_{\Delta t}(x|y) \, p(y, t) \\ \{y = x - \epsilon\} = & \int_{\mathbb{R}^n} \mathrm{d}\epsilon q_{\Delta t}(x|x - \epsilon) \, p(x - \epsilon, t) \\ = & \int_{\mathbb{R}^n} \mathrm{d}\epsilon q_{\Delta t}((x - \epsilon) + \epsilon|x - \epsilon) \, p(x - \epsilon, t) \\ \{\omega := \cdots\} = & \int_{\mathbb{R}^n} \mathrm{d}\epsilon \omega(x - \epsilon, \epsilon) \, p(x - \epsilon, t). \end{aligned}$$

And for the second term,

$$\begin{aligned} & \int_{\mathbb{R}^n} \mathrm{d}y q_{\Delta t}(y|x) p(x, t) \\ \{y = x - \epsilon\} = & \int_{\mathbb{R}^n} \mathrm{d}(-\epsilon) \, q_{\Delta t}(x - \epsilon|x) \, p(x, t) \\ \{-\epsilon \rightarrow \epsilon\} = & \int_{\mathbb{R}^n} \mathrm{d}\epsilon q(x + \epsilon|x) \, p(x, t) \\ \{\omega := \cdots\} = & \int_{\mathbb{R}^n} \mathrm{d}\epsilon \omega(x, \epsilon) \, p(x, t). \end{aligned}$$

Altogether, we have

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

Now, since $q_{\Delta t}$ and p are smooth, we can Taylor expand the first term, and find

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

All together, we get

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

By denoting

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

we arrive at

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

This is called the **Kramers–Moyal expansion**.

Recalling that $\omega(x, \epsilon) = q_{\mathrm{d}t}(x + \epsilon|x)$, we have

$$M^{\alpha^1 \cdots \alpha^k}(x) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon \, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k}) \, q_{\mathrm{d}t}(x + \epsilon|x) = \int_{\mathbb{R}^n} \mathrm{d}\epsilon \, (\epsilon^{\alpha_1} \cdots \epsilon^{\alpha_k}) \, \omega(x, \epsilon)$$

so $M^{\alpha^1 \cdots \alpha^k}(x)$ is recognized as the k-order moment of ϵ sampled from transition density $q_{\Delta t}(x + \epsilon|x)$ (regarding $q_{\Delta t}(x + \epsilon|x)$ as a distribution $Q_{\Delta t}(\epsilon)$).

3.2. Langevin Dynamics that Satisfies Detailed Balance Is Conservative

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, the transition density of **Langevin dynamics**, $q_{\mathrm{d}t}(x'|x)$, is a normal distribution of $x'-x$ with mean value $\mu(x)\,\mathrm{d}t$ and variance $2\Sigma(x)\mathrm{d}t$. Thus, $M^{\alpha}(x)=\mu^{\alpha}(x)\,\mathrm{d}t$, $M^{\alpha\beta}(x)=2\Sigma^{\alpha\beta}(x)\,\mathrm{d}t$, and higher orders are of $o(\mathrm{d}t)$. The Kramers-Moyal expansion gives

$$\frac{\partial p}{\partial t}(x,t)=-\nabla_{\alpha}(\mu^{\alpha}(x)\,p(x,t))+\nabla_{\alpha}\nabla_{\beta}(\Sigma^{\alpha\beta}(x)\,p(x,t)), \tag{15}$$

which is the **Fokker-Planck equation**.

As a special case of master equation, we may wonder when Fokker-Planck equation will satisfy detailed balance condition? Directly from the form of transition density, we find that if there is a stationary distribution Π such that Fokker-Planck equation satisfies detailed balance condition, then we must have [9]

$$\mu^{\alpha}(x)=\Sigma^{\alpha\beta}(x)\nabla_{\beta}\left[\ln\pi(x)-\frac{1}{2}\operatorname{tr}\ln\Sigma(x)\right]. \tag{16}$$

This indicates that, to satisfy detailed balance condition, μ shall be conservative.[10]

4. Maximum-Entropy Principle

4.1. Conventions in This Section

Follow the conventions in section 2.

4.2. Maximum-Entropy Principle Shall Minimize Relative Entropy

As discussed in section 1, Shannon entropy is not well-defined for continuous random variable, while the relative entropy is proper for both discrete and continuous random variables. Comparing with Shannon entropy, relative entropy needs an extra distribution, which describes the prior knowledge. It then characterizes the relative uncertainty (surprise) of a distribution to the distribution of prior knowledge. When the prior knowledge is unbiased and $|\mathcal{X}|:=\int_{\mathcal{X}}\mathrm{d}x\,1<+\infty$, the negative relative entropy reduces to Shannon entropy. So, maximum-entropy principle shall minimize relative entropy.

[9] Suppose there is a stationary distribution π such that $q_{\mathrm{d}t}(x+\epsilon|x)\,\pi(x)=q_{\mathrm{d}t}(x|x+\epsilon)\pi(x+\epsilon)$. Since $q_{\mathrm{d}t}(x+\epsilon|x)$ obeys normal distribution $\mathcal{N}(\mu(x)\mathrm{d}t, 2\Sigma(x)\mathrm{d}t)$ on ϵ, the the relation comes to be

$$\begin{aligned}&\frac{1}{\sqrt{(4\pi)^n\det[\Sigma(x)]}}\exp\left(-\frac{1}{4\mathrm{d}t}(\epsilon-\mu(x)\mathrm{d}t)\cdot\Sigma^{-1}(x)\cdot(\epsilon-\mu(x)\mathrm{d}t)\right)\pi(x)\\ ={}&\frac{1}{\sqrt{(4\pi)^n\det[\Sigma(x+\epsilon)]}}\exp\left(-\frac{1}{4\mathrm{d}t}(-\epsilon-\mu(x+\epsilon)\mathrm{d}t)\cdot\Sigma^{-1}(x+\epsilon)\cdot(-\epsilon-\mu(x+\epsilon)\mathrm{d}t)\right)\pi(x+\epsilon).\end{aligned}$$

Notice that

$$\begin{aligned}\ln\det[\Sigma(x+\epsilon)]&=\ln\det[\Sigma(x)+(\epsilon\cdot\nabla)\Sigma(x)]\\&=\ln\det[\Sigma(x)]+\ln\det[1+(\epsilon\cdot\nabla)(\Sigma^{-1}(x)\cdot\Sigma(x))]\\&=\ln\det[\Sigma(x)]+\ln\{1+\operatorname{tr}[(\epsilon\cdot\nabla)(\Sigma^{-1}(x)\cdot\Sigma(x))]\}\\&=n\det[\Sigma(x)]+\operatorname{tr}[(\epsilon\cdot\nabla)(\Sigma^{-1}(x)\cdot\Sigma(x))]\\&=\ln\det[\Sigma(x)]+\epsilon\cdot\nabla\operatorname{tr}\ln\Sigma.\end{aligned}$$

The typical order of ϵ is $\mathcal{O}\left(\sqrt{\Sigma(x)\,\mathrm{d}t}\right)$, or say, $\mu(x)\mathrm{d}t=\mathcal{O}(\epsilon^2\,\mu(x)/\Sigma(x))$. If $\mu(x)=\mathcal{O}(\Sigma(x))$, then we have $\mu(x)\,\mathrm{d}t=(\epsilon^2)$. So, we have

$$-\frac{1}{4\mathrm{d}t}(-\epsilon-\mu(x+\epsilon)\mathrm{d}t)\cdot\Sigma^{-1}(x+\epsilon)\cdot(-\epsilon-\mu(x+\epsilon)\mathrm{d}t)=-\frac{1}{4\mathrm{d}t}(-\epsilon-\mu(x)\mathrm{d}t)\cdot\Sigma^{-1}(x)\cdot(-\epsilon-\mu(x)\mathrm{d}t)+o(\epsilon^2).$$

Altogether, expanding the first formula on both sides by ϵ to the lowest order gives

$$\mu(x)=\Sigma(x)\cdot\nabla\left[\ln\pi(x)-\frac{1}{2}\operatorname{tr}\ln\Sigma(x)\right].$$

[10] Recall that Σ is symmetric thus can be diagonalized, the $\Sigma^{\alpha\beta}(x)$ factor can be then be absorbed by a redefinition of x and $\mu(x)$, so that vector field μ is the gradient of a scalar function, that is, being conservative.

Given a distribution Q that describes the prior knowledge of random variable X, the basic problem is to find a distribution P of X such that the relative entropy $H(P,Q)$ is minimized under a set of restrictions $\{\mathbb{E}_P[f_\alpha] = \bar{f}_\alpha | \alpha = 1, \ldots, m, f_\alpha: \mathcal{X} \to \mathbb{R}\}$. The notation $\mathbb{E}_P[\cdots] := \int_{\mathcal{X}} \mathrm{d}x p(x) \cdots$ represents expectation under P; and the function f_α is called **observable** and the value $\bar{f}_\alpha$ is called an **observation**. Thus, P is the distribution that is closest to the prior knowledge with the restrictions fulfilled.

To solve this problem, we use variational principle with Lagrangian multipliers. There are two kinds of constraints. One from the restrictions $\mathbb{E}_P[f_\alpha] = \bar{f}_\alpha$ for each α; and the other from normalization $\int_{\mathcal{X}} \mathrm{d}x p(x) = 1$. Recall that the relative entropy $H(P,Q) := \int_{\mathcal{X}} \mathrm{d}x p(x) \ln(p(x)/q(x))$. Altogether, the loss functional becomes

$$L(p, \lambda, \mu) := \int_{\mathcal{X}} \mathrm{d}x p(x) \ln \frac{p(x)}{q(x)} + \lambda^\alpha \left(\int_{\mathcal{X}} \mathrm{d}x p(x) f_\alpha(x) - \bar{f}_\alpha \right) + \mu \left(\int_{\mathcal{X}} \mathrm{d}x p(x) - 1 \right). \tag{17}$$

So, we have (L is a functional of p, thus use δ instead of ∂ for p),

$$\begin{aligned} \frac{\delta L}{\delta p(x)}(p, \lambda, \mu) &= \ln p(x) + 1 - \ln q(x) + \lambda^\alpha f_\alpha(x) + \mu; \\ \frac{\partial L}{\partial \lambda^\alpha}(p, \lambda, \mu) &= \int_{\mathcal{X}} \mathrm{d}x p(x) f_\alpha(x) - \bar{f}_\alpha; \\ \frac{\partial L}{\partial \mu}(p, \lambda, \mu) &= \int_{\mathcal{X}} \mathrm{d}x p(x) - 1. \end{aligned}$$

These equations shall vanish on the extremum. If $(p_\star, \lambda_\star, \mu_\star)$ is an extremum, then

$$\frac{\partial \ln Z}{\partial \lambda^\alpha}(\lambda_\star) + \bar{f}_\alpha = 0 \tag{18}$$

for each $\alpha = 1, \ldots, m$, where

$$Z(\lambda) := \int_{\mathcal{X}} \mathrm{d}x q(x) \exp(-\lambda^\alpha f_\alpha(x)); \tag{19}$$

and

$$p_\star(x) = p(x, \lambda_\star), \tag{20}$$

where

$$p(x, \lambda) := q(x) \exp(-\lambda^\alpha f_\alpha(x)) / Z(\lambda). \tag{21}$$

The $\mu_\star$ has been included in the Z.

4.3. Prior Knowledge Furnishes Free Theory or Regulator

Compared with the maximum-entropy principle derived from maximizing Shannon entropy, we get an extra factor $q(x)$ in $p(x, \lambda)$. This factor plays the role of prior knowledge.

In physics, this prior knowledge can be viewed as free theory, a theory without interactions. Indeed, interaction shall be given by the restrictions, the expectations of observables. It is the factor $\exp(-\lambda^\alpha f_\alpha(x))$ in $p(x, \lambda)$. The λ plays the role of couplings. This indicates that $q(x)$ shall be the free theory.

In machine learning, it acts as regulator, a pre-determined term employed for regulating the value of x.

4.4. When Is $\lambda_\star$ Solvable? (TODO)

Even though it is hard to guarantee the equation 18 solvable, we have some results for the case when $\bar{f} \approx \mathbb{E}_Q[f]$. That is, the perturbative case.

To guarantee that perturbative solution exists for equation 18, we have to ensure that the Jacobian $\partial^2 \ln Z / \partial \lambda^\alpha \partial \lambda^\beta$ is not degenerate at $\lambda = 0$. With a series of direct calculation, we find

$$\frac{\partial^2 \ln Z}{\partial \lambda^\alpha \partial \lambda^\beta}(0) = \mathrm{Cov}_q(f_\alpha, f_\beta), \tag{22}$$

the covariance matrix of f under distribution q.

5. 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.

5.1. Conventions in This Section

Follow the conventions in section 2. 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.

5.2. Data Fitting Is Equivalent to Least-Action Principle

Let $P(\theta)$ represent a parametrized distribution of X, and $\hat{P}$ a distribution of X that represents prior knowledge as in the case of maximum-entropy principle. Let $S(x, \theta) := -\ln (p(x, \theta) / \hat{p}(x)) - \ln Z(\theta)$ with $Z(\theta)$ to be determined. Density $\hat{p}$ is essential for defining S, since $\ln p(x, \theta)$ is not well-defined (section 1.3). Then, we can re-formulate $p(x, \theta)$ as

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

and since $\int_{\mathcal{X}} \mathrm{d}x \, p(x, \theta) = 1$,

$$Z(\theta) = \int_{\mathcal{X}} \mathrm{d}x \, \hat{p}(x) \exp(-S(x, \theta)). \tag{24}$$

As a generic form of a parameterized distribution, it can be used to fit raw data that obeys an empirical distribution Q, by adjusting parameter θ. To do so, we minimize the relative entropy between Q and $P(\theta)$, which is defined as $H(Q, P(\theta)) := \int_{\mathcal{X}} \mathrm{d}x \, q(x) \ln (q(x) / p(x, \theta))$. Plugging equation (23) into $H(Q, P(\theta))$, we have

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

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

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

We can find the $\theta_\star := \operatorname{argmin} L$ by iteratively updating θ along the direction $-\partial L / \partial \theta$. With a series of direct calculus,[11] we find

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

[11] Directly, we have

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

Since $Z(\theta) := \int \mathrm{d}x \, \hat{p}(x) \exp(-S(x, \theta))$, we find

$$\frac{\partial Z}{\partial \theta^\alpha}(\theta) = -\int \mathrm{d}x \, \hat{p}(x) \exp(-S(x, \theta)) \frac{\partial S}{\partial \theta^\alpha}(x, \theta),$$

thus

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

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

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

$$\mathbb{E}_{P(\theta_\star)}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right] = \mathbb{E}_Q\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right]. \tag{26}$$

Notice that L is equivalent to another loss L_{LA} where

$$L_{\mathrm{LA}}(\theta) := \mathbb{E}_Q[S(\cdot,\theta)] - \mathbb{E}_{P(\theta)}[S(\cdot,\theta)]. \tag{27}$$

The expectation $\mathbb{E}_{P(\theta)}$ is computed by Monte-Carlo method. We sample data points from $P(\theta)$ with fixed θ, and compute the mean value of $S(\cdot,\theta)$ on these data points. *The derivative of θ on this expectation is taken on the $S(\cdot,\theta)$ instead of on the data points.* In this way, L_{LA} is equivalent to L.

It can be read from this equation that minimizing L_{LA} is to decrease the $S(\cdot,\theta)$ at data points (the first term) while increase it at the points away from data (the second term). As figure 2 illustrates, this way of optimization will site a real world datum onto a local minimum of $S(\cdot,\theta)$, in statistical sense. In this way, the $S(\cdot,\theta)$ is recognized as a parameterized action. It thus describes the dynamics of an entity. This entity may be of physics, like particles. But it can also be words, genes, flock of birds, and so on. For example, we can find out how words "interact" with each other.

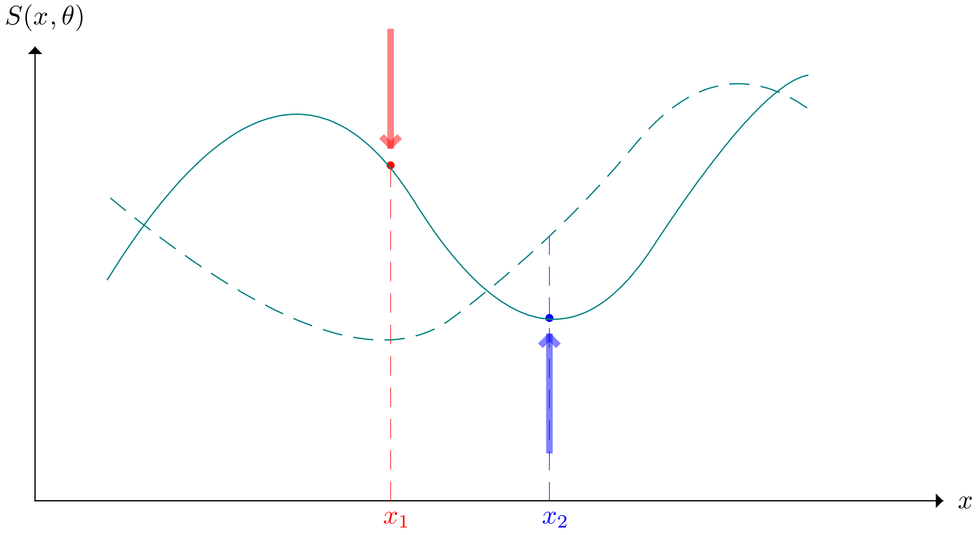

Figure 2. This figure illustrate how $\min_\theta L_{\mathrm{LA}}(\theta)$ will site a real world datum onto a local minimum of $S(\cdot,\theta)$. The green curve represents the current not-yet-optimized $S(\cdot,\theta)$. The x_1 (red point) is a real world datum while x_2 (blue point), which is currently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

5.3. Example: Extract Dynamics from Raw Data

We are to apply the previous discussion to extract dynamics from the raw data of a physical system. To describe the system, we need a configuration like $x(t)$. So, the raw data is a set $\{(x_k(1),\ldots,x_k(T))|k=1,\ldots,D\}$ where time is discretized as $(1,\ldots,T)$ and the data size is D. Thus, each datum is a movie of the physical system, frame by frame. These raw data are obtained by experiments and measurements (with measurement errors).

As a physical system, the $\hat{p}$ that represents free theory shall be Gaussian. It may be

$$\hat{p}(x) \propto \exp\left\{-\frac{1}{2}\sum_{t=1}^{T-1} [x(t+1)-x(t)]^2\right\},$$

indicating a kinetic term.

The action $S(x,\theta)$ is given by some ansatz. First, we may suppose that the action is local. That is, there is a Lagrangian $L(x,t,\theta)$ such that $S(x,\theta)=\sum_{t=1}^{T} L(x(t),t,\theta)$. Next, we may suppose that there exist some symmetries about the physical system, such as autonomous and parity symmetry, which means $L(x,t,\theta)=\sum_{n=1}^{+\infty}\theta_n x^{2n}$ when x is 1-dimensional. These symmetries will further restrict the possible form of the action. Finally, we can write down a most generic form of action that satisfies all the ansatz. Neural network and symbolic regression may help you write down this most generic form. Then, we find the best fit $\theta_\star$ by equation 25. The action $S(x,\theta_\star)$ describes the dynamics extracted from the raw data.[12]

5.4. Is There an Action for a Dynamical System?

The local minima of L_{LA} (the loss function used for finding action) can be realized as the patterns in the dataset. So, *we can find an action for a physical system if and only if there are a finite number of patterns in the system* (that is, in the raw data). When a physical system is chaotic, there will be an infinite number of patterns in it. In fact, a chaotic system can be seen as an advanced generator of pseudo-random numbers, being avoid of any pattern. TODO

Consider the one-dimensional hamonic oscillator

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}(t)+\omega^2 x(t)=0.$$

The $L(x,t)=(1/2)\,\omega^2 x^2$. The general solution is a linear combination of two bases $\sin(\omega t)$ and $\cos(\omega t)$, thus

$$x(t)=A\sin(wt)+B\cos(\omega t).$$

For simplicity, let us first consider $x(t)=A\sin(\omega t)$, thus $x(0)=0$ and $\dot{x}(0)=A\omega$ It has one pattern which is $\sin(\omega t)$ but gauged by A. It gives

$$S(x)=\frac{\omega^2 A^2}{2}\cos^2(\omega t)-\frac{\omega^2 A^2}{2}\sin^2(\omega t)=\frac{\dot{x}^2(0)}{2}\cos(2\omega t).$$

5.5. Example: Actions in Machine Learning (TODO)

In section 5.2, we have shown that any density function can be re-formulated by action. Usually, the goal of a supervised machine learning task is to fit a density function that predicts the target. For example, given an image, we are to compute the conditional density function for the class of the image such as being a cat or a dog. Let x denote the input (like images) and y the target, which can be discrete (like classes) or continuous (like person's height), then the conditional density function is usually given by a model $f(x,\theta)$ parameterized by θ, as

$$p(y|x,\theta)=\mathcal{P}(y,f(x,\theta)).$$

For a categorical classification task, $y\in\{1,\ldots,M\}$ and $z\in\mathbb{R}^M$ for some M, and $\mathcal{P}$ is defined by

$$\mathcal{P}_{\mathrm{clf}}(y,z):=\frac{\exp(z^y)}{\sum_{\alpha=1}^{M}\exp(z^\alpha)}.$$

And for regression task, $y,z\in\mathbb{R}^M$ for some M, and $\mathcal{P}$ is defined by

$$\mathcal{P}_{\mathrm{rg}}(y,z):=\exp\left(-\sum_{\alpha=1}^{M}\frac{(y^\alpha-z^\alpha)^2}{2}\right).$$

[12] An experiment on general oscillators can be found in the oscillators/Oscillator.ipynb.

Thus, we have an action

$$S(y;x,\theta) := -\ln p(y|x,\theta) = -\ln \mathcal{P}(y, f(x,\theta)),$$

which is the loss per sample in machine learning.[13] Remark that input x serves as a parameter of action S, and y is unique the argument of action (so we use semicolon instead of comma).

Assume that datum (x,y) is sampled from a dataset described by distribution Q, thus the total loss of least-action becomes (Q_X for the marginal distribution of X, and $P(x,\theta)$ for the conditional distribution of $p(y|x,\theta)$, thus we can sample from it)

$$L_{\text{LA}}(\theta) = \mathbb{E}_{x\sim Q_X, y\sim P(x,\theta)}[\ln \mathcal{P}(y, f(x,\theta))] - \mathbb{E}_{(x,y)\sim Q}[\ln \mathcal{P}(y, f(x,\theta))].$$

The last term is the usual loss function in machine learning. For example, in the classification task, it is cross-entropy, and in regression task, it is usually the mean squared error.

The first term is new for machine learning. To compute it, we first sample a datum (x, y_0) from Q and only keep the x, which indicates the $x \sim Q_X$. Then, compute $f(x,\theta)$ and sample a new y from $P(x,\theta)$. For classification task, y is sampled from the categorical distribution with probability $\exp(f^y(x;\theta)) / \sum_{\alpha=1}^{M} \exp(f^\alpha(x,\theta))$, and for regression task, from a normal distribution with mean $f(x,\theta)$ and unit variance. Using this y, together with the x, $\ln \mathcal{P}(y, f(x,\theta))$ is calculated. For classification task, it is $f^y(x,\theta) - \text{lnSumExp}(f(x,\theta))$, where $\text{lnSumExp}(x) := \ln(\sum_\alpha \exp(x^\alpha))$; and for regression task, it is $-\sum_{\alpha=1}^{M} (y^\alpha - f^\alpha(x,\theta))^2/2$.

When we use deep neural network to express the model $f(x,\theta)$, TODO

There is also unsupervised learning tasks. For example, mask some part of the input image and predict what has been masked, or embedding the input image into a latent space. TODO

5.6. Maximum-Entropy and Least-Action Are Saddle Point of a Functional

In fact, equations (23), (24), and (26) can be regarded as an extremum of the functional (recall that $\hat{P}$ is the distribution of prior knowledge and Q of dataset)

$$V(p,\theta,\mu) := H[P,\hat{P}] + (\mathbb{E}_P[S(\cdot,\theta)] - \mathbb{E}_Q[S(\cdot,\theta)]) + \mu(\mathbb{E}_P[1] - 1),$$

or explicitly

$$V(p,\theta,\mu) = \int_{\mathcal{X}} \mathrm{d}x p(x) \ln\frac{p(x)}{\hat{p}(x)} + \left(\int_{\mathcal{X}} \mathrm{d}x p(x) S(x,\theta) - \int_{\mathcal{X}} \mathrm{d}x q(x) S(x,\theta)\right) + \mu\left(\int_{\mathcal{X}} \mathrm{d}x p(x) - 1\right). \tag{28}$$

Indeed, variance on p gives equation (23).[14] Together with the partial derivative on μ, we get equation (24). Finally, partial derivative on θ directly gives equation (26). Interestingly, the second term is just the $-L_{\text{LA}}(\theta)$ in equation (27). So, the extremum is in fact a saddle point, as

$$(p_\star, \theta_\star, \mu_\star) = \min_{p,\mu} \max_{\theta} V(p,\theta,\mu). \tag{29}$$

By tuning p, the $\min_{p,\mu}$ minimizes the relative entropy between P and Q and the expectation of action $\mathbb{E}_P[S(\cdot,\theta)]$, which in turn relates the density function p with the action $S(\cdot,\theta)$. And by tuning θ, the $\max_\theta$ sites real data onto the action's local minima. So, we find that maximum-entropy principle and least-action principle are saddle point of a functional V.

5.7. Structures in Nature Arise from Maximum-Entropy (TODO)

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 does these structures arise in nature?

[13] So, machine learning models can be seen as a complicated version of Ising model, where the y in machine learning corresponding to the spins in Ising model and θ to the temperature.

[14] Explicitly, we have

$$\frac{\delta V}{\delta p(x)}(p,\theta,\mu) = \ln p(x) + 1 - \ln \hat{p}(x) + S(x,\theta) + \mu = 0,$$

which has solution

$$p(x) \propto \hat{p}(x) \exp(-S(x,\theta)).$$

The vascular system is fine-tuned so as to minimize the frictional loss. The chromosome that determines the phenotype of an organism is also fine-tuned such that the probability of survival is maximal. These examples indicate that structure appears in optimizing an objective.

So, let random variable X characterize the configuration, such as the sequence of bases along chromosome. There is an action that reflects the interaction of bases and the environment. The one that survives has the most "coherent" chromosome that minimizes the action. The "survival" distribution is given by the action. This distribution has many local maxima. The maximal local maxima represents the creature that has the highest adaptation, maybe human.

The action would be very complex. But, inversely, given the real world data of chromosome, the action can be revealed by parameterized function and least-action principle. To do so, we first travel to a closed island, such as Galápagos Islands, which forms a closed system. Then, we collect the chromosomes of all creatures living on the island. Since different creatures have different lengths of chromosome, we have to unify the coding of chromosomes. This furnishes the alphabet $\mathcal{X}$, and the collection of chromosomes characterizes the distribution of real world data, Q. Let $S(x, \theta)$ a parameterized function, with parameters θ. The least-action principle gives the best fit $\theta_\star$ by minimizing the L_{LA}. During the minimization, we have to sample from $P(\theta)$, where the density function $p(x, \theta) \propto \exp(-S(x, \theta))$. The strategy is using a transition rate that satisfies the detailed balance condition. This transition rate minics the evolution. The disconnectivity of transition rate may reflect gene isolation.