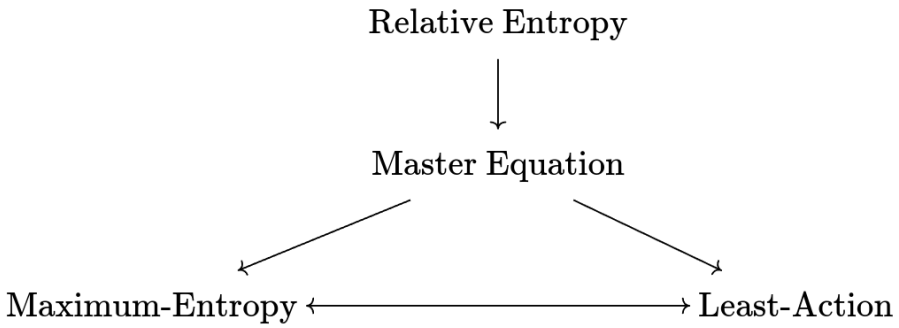

Figure 1. Scheme of sections. The solid arrow represents relation.

1. Relative Entropy

1.1. A Short Review of Probability

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

The set of all possible values of a random variable is called the **alphabet**.[1] And for each value in the alphabet, we assign a *positive* value called **density** if the alphabet is of continuum (continuous random variable), or **mass** otherwise (discrete random variable).[2] We use **distribution** for not only the mass or density on the alphabet, but also a sampler that can sample an ensemble of values of the random variable that converges to the mass or density when the number of sample tends to infinity. For example, we say X is a random variable with alphabet $\mathcal{X}$ and distribution P.

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

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

1.2. Shannon Entropy Is Plausible for Discrete 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.

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.

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.

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,

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

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 $\partial_1 h(1, 1) \int \mathrm{d}x p(x) \ln(p(x)/p(x)) + C = 0$. 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 Generic Dynamics of Markov Chain

The generic dynamics of a Markov chain can be characterized by its **transition density** $q_{t \to t'}(y|x)$ which describes the density that transits from x at time t to y at time t'. Since the underlying dynamics which determines $q_{t \to t'}$ is usually autonomous, we can suppose that $q_{t \to t'}$ depends only on the difference $\Delta t := t' - t$. This will greatly reduce the complexity while covering most of the important situations. So, throughout this note, we use $q_{\Delta t}$ instead of $q_{t \to t'}$.

During a temporal unit Δt, the change of density at $X = x$ equals to the total density that transits into x subtracting the total density that transits out of x. That is,

$$p(x, t + \Delta t) - p(x, t) = \int_{\mathcal{X}} \mathrm{d}y q_{\Delta t}(x|y) p(y, t) - \int_{\mathcal{X}} \mathrm{d}y q_{\Delta t}(y|x) p(x, t). \tag{1}$$

It is called the **discrete time master equation**.

What are the requirements for transition density? The last term of master equation 1 can be re-written as $p(x,t)\times\int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(y|x)$. If $q_{\Delta t}$ is normalized by $\int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(y|x)=1$ for any $x\in\mathcal{X}$, then we have

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

It means the density at x equals to the density transited into x during the time interval Δt (this is a little weird). We have $p(x,t)>0$ for all $x\in\mathcal{X}$, but have to guarantee that $p(x,t+\Delta t)>0$, for sharing the same alphabet $\mathcal{X}$. Basically, this requires $q_{\Delta t}(x|y)$ to be non-negative for all x, $y\in\mathcal{X}$, since, otherwise, we can construct a $p(y,t)$ with large value on the place where $q_{\Delta}(x|y)<0$ and tiny value on the rest, so that $p(x,t+\Delta t)<0$. This is natural since $q_{\Delta}(x|y)$ describes the transited density, and negative transited density cannot be "physical". It further requires that for each $x\in\mathcal{X}$, there is density transited into x. Precisely, for any x there is a measurable subset of y such that $q_{\Delta t}(x|y)>0$. Combined with the non-negativity, it directly implies $p(x,t+\Delta t)>0$. As a summary, we require the transition density $q_{\Delta t}(x|y)$ to be normalized on x, to be non-negative, and to be that there is measurable density transited into x (or equivalently, $\int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(x|y)>0$).

If we apply $q_{\Delta t}$ twice to $p(x,t)$, we get $p(x,t+2\Delta t)$:

$$\begin{aligned}
& p(x,t+2\Delta t)\\
\{\text{master equation}\} &= \int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(x|y)p(y,t+\Delta t)\\
\{\text{master equation}\} &= \int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(x|y)\int_{\mathcal{X}}\mathrm{d}z q_{\Delta t}(y|z)p(z,t)\\
&= \int_{\mathcal{X}}\mathrm{d}z\left[\int_{\mathcal{X}}\mathrm{d}y\, q_{\Delta t}(x|y)\, q_{\Delta t}(y|z)\right]p(z,t)\\
\{\text{definition of } q_{2\Delta t}\} &= \int_{\mathcal{X}}\mathrm{d}z q_{2\Delta t}(x|z)\, p(z,t),
\end{aligned}$$

thus, $q_{2\Delta t}(x|z)=\int_{\mathcal{X}}\mathrm{d}y\, q_{\Delta t}(x|y)\, q_{\Delta t}(y|z)$. Repeat this process, we will arrive at

$$q_{(n+1)\,\Delta t}(x|z)=\int_{\mathcal{X}}\mathrm{d}y_1\cdots\int\mathrm{d}y_n\, q_{\Delta t}(x|y_1)\cdots q_{\Delta t}(y_n|z). \tag{2}$$

This is a little like path integral: we sum over all the paths from z to x.

If $q_{\Delta t}$ is smooth on Δt, then we have the linear expansion $q_{\Delta t}(x|y)=\chi(x,y)+r(x,y)\,\Delta t+o(\Delta t)$ where q_0 and r are well-defined functions (at least well-defined generalized functions). When $\Delta t=0$, there is no transition at all, and $p(x,t)=\int_{\mathcal{X}}\mathrm{d}y\chi(x,y)p(y,t)$. The density function p is arbitrary, implying that $\chi(x,y)=\delta(x-y)$. Plugging all these back to equation 1 and taking the limit $\Delta t\to 0$, we get the **continuous time master equation**:

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

The r is called the **transition rate**, characterizing the speed of transition. The normalization $\int_{\mathcal{X}}\mathrm{d}x\, q_{\Delta t}(x|y)=1$ demands that $\int_{\mathcal{X}}\mathrm{d}x r(x,y)=0$. And the non-negativity $q_{\Delta t}(x|y)\geqslant 0$ demands that $r(x,y)\geqslant 0$ when $x\neq y$. The last requirement of transition density demands that $\int_{\mathcal{X}}\mathrm{d}y r(x,y)>-1$, which is a little weird.

2.3. Detailed Balance Provides Stationary Distribution

Let Π a stationary solution of master equation 3. Then, Π satisfies $\int_{\mathcal{X}}\mathrm{d}y\,[r(x,y)\,\pi(y)-r(y,x)\pi(x)]=0$. But, this condition is too weak to be used. A more useful condition, which is stronger than this, is that the integrand vanishes everywhere:

$$r(x,y)\,\pi(y)=r(y,x)\pi(x), \tag{4}$$

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

Given a transition rate, we wonder if there exists a density function such that detailed balance 4 holds. This problem is very complicated. In many applications, we consider the inverse: given a density function, if there exists a transition rate such that detailed balance holds. This inverse problem is much simpler, and a proper transition rate can be constructed out of the density function (such as in Metropolis-Hastings algorithm).

2.4. Detailed Balance 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

$$H(P(t),\Pi)=\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)\ln\frac{p(x,t)}{\pi(x)}. \tag{5}$$

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[5]

$$\frac{\mathrm{d}}{\mathrm{d}t}H(P(t),\Pi)=-\frac{1}{2}\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y\,r(y,x)\,\pi(x)\left[\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right]\left[\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right]. \tag{6}$$

5. 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 $r(x,y)=r(y,x)\,\pi(x)/\pi(y)$, as

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

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

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

By adding the two previous results together, we find

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

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

We are to check the sign of the integrand. The $r(y,x)$ is negative only when $x=y$, on which the integrand vanishes. Thus, $r(y,x)$ can be treated as non-negative, so is the $r(y,x)\,\pi(x)$ (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 $r(y,x)=0$ or $p(x,t)/\pi(x)=p(y,t)/\pi(y)$. If $r(y,x)>0$ 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 $r(y,x)=0$ 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**. In this situation, $(\mathrm{d}/\mathrm{d}t)\,H(P(t),\Pi)=0$ implies $p(x,t)/\pi(x)=p(z,t)/\pi(z)$.[6] So, by repeating the previous discussion on the case "$r(y,x)>0$ 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 seperates $\mathcal{X}$ 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 seperated 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 such that 1) every two elements in $\mathcal{X}$ are connected and that 2) the detailed balance 4 holds for Π, 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).

2.5. Temporal Smoothness of Transition Density Is Necessary to Ensure Relaxation

The temporal smooth structure, however, cannot be avoided. Indeed, the smoothness of transition density on time and thus the smoothness of $p(x,t)$ on t is essential for the monotonic reduction of relative entropy, which is the essential end of our discussion.[7]

6. 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 $z=y_n$.

7. You may wonder if the temporal smoothness implies the continuum of alphabet. Explicitly, if $p(x,t)$ is smooth on t, then does the value of x have to be continuous? The answer is no. For example, consider an alphabet $\mathcal{X}=\{0,1\}$; the $p(1,t)$ is given by $\sigma(\zeta(t))$ where σ denotes the sigmoid function and $\zeta(t)$ is smooth on t, thus $p(0,t)=1-\sigma(\zeta(t))$. In this example, $p(x,t)$ is smooth on t but the random variable is discrete.

To see this clearly, let us examine $H(P(t+\Delta t),\Pi)-H(P(t),\Pi)$ when Δt is not an infinitesimal. By definition,

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

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

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

The first line is recognized as $H(P(t+\Delta t),P(t))$, which is non-negative. Following the same steps in section 2.4 (but using discrete time master equation 1 instead), the second line reduces to

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

which is non-positive. The sign of the final result can be arbitrary. Indeed, the first line is determined by the difference between $p(\cdot,t+\Delta t)$ and $p(\cdot,t)$, while the second line is determined by the difference between $p(\cdot,t)$ and π. They are intrinsically different, thus mutually independent. So, we conclude that the smoothness of $q_{\Delta t}$ on Δt is essential for the guarantee of the monotonic reduce of relative entropy between $p(\cdot,t)$ and π, thus its relaxation.

3. Kramers-Moyal Expansion and Langevin Dynamics

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 master equation 1 becomes

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

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,

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

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

$$p(x,t+\Delta t) - p(x,t) = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \left(\frac{\partial}{\partial x^{\alpha_1}} \cdots \frac{\partial}{\partial x^{\alpha_k}} \right) [M^{\alpha^1 \cdots \alpha^k}(x)\, p(x,t)]. \tag{7}$$

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

$$\frac{\partial p}{\partial t}(x,t) = -\nabla_\alpha(\mu^\alpha(x)\, p(x,t)) + \nabla_\alpha \nabla_\beta(\Sigma^{\alpha\beta}(x)\, p(x,t)), \tag{8}$$

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

As a special case of master equation, we may wonder when Fokker-Planck equation will satisfy detailed balance? 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, then we must have [8]

$$\mu^\alpha(x) = \Sigma^{\alpha\beta}(x)\nabla_\beta\left[\ln\pi(x) - \frac{1}{2}\operatorname{tr}\ln\Sigma(x)\right]. \tag{9}$$

This indicates that, to satisfy detailed balance, μ shall be conservative.[9]

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.

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 $\{\mathbb{E}_P[f_\alpha] = \bar{f}_\alpha | \alpha = 1,\dots,m,\ f_\alpha\colon \mathcal{X}\to\mathbb{R}\}$. 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.

8. 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 $\mathcal{O}\big(\sqrt{\Sigma(x)\,\mathrm{d}t}\big)$, 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

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

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

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

$$L(p,\lambda,\mu) := \int_{\mathcal{X}} \mathrm{d}x\, p(x) \ln\frac{p(x)}{q(x)} + \lambda^\alpha \left(\int_{\mathcal{X}} \mathrm{d}x\, p(x) f_\alpha(x) - \bar{f}_\alpha \right) + \mu \left(\int_{\mathcal{X}} \mathrm{d}x\, p(x) - 1 \right). \tag{10}$$

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

$$\frac{\partial \ln Z}{\partial \lambda^\alpha}(\lambda_\star) + \bar{f}_\alpha = 0 \tag{11}$$

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

$$Z(\lambda) := \int_{\mathcal{X}} \mathrm{d}x\, q(x) \exp(-\lambda^\alpha f_\alpha(x)); \tag{12}$$

and

$$p_\star(x) = p(x, \lambda_\star), \tag{13}$$

where

$$p(x,\lambda) := q(x) \exp(-\lambda^\alpha f_\alpha(x)) / Z(\lambda). \tag{14}$$

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 11 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 11, 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

$$\frac{\partial^2 \ln Z}{\partial \lambda^\alpha \partial \lambda^\beta}(0) = \mathrm{Cov}_q(f_\alpha, f_\beta), \tag{15}$$

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 $S(x,\theta) := -\ln\left(p(x,\theta)\,/\,\hat{p}(x)\right) - \ln Z(\theta)$ 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

$$p(x,\theta) = \hat{p}(x)\exp(-S(x,\theta))\,/\,Z(\theta), \tag{16}$$

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

$$Z(\theta) = \int_{\mathcal{X}} \mathrm{d}x\, \hat{p}(x)\exp(-S(x,\theta)). \tag{17}$$

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 $H(Q,P(\theta)) := \int_{\mathcal{X}} \mathrm{d}x\, q(x)\ln\left(q(x)\,/\,p(x,\theta)\right)$. Plugging equation (16) 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,[10] we find

$$\frac{\partial L}{\partial\theta^\alpha}(\theta) = \mathbb{E}_Q\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta)\right] - \mathbb{E}_{P(\theta)}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta)\right]. \tag{18}$$

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

$$\mathbb{E}_{P(\theta_\star)}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right] = \mathbb{E}_Q\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right]. \tag{19}$$

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

$$L_{\mathrm{LA}}(\theta) := \mathbb{E}_Q[S(\cdot,\theta)] - \mathbb{E}_{P(\theta)}[S(\cdot,\theta)]. \tag{20}$$

10. 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 $-\mathbb{E}_{P(\theta)}[(\partial S\,/\,\partial\theta^\alpha)(\cdot,\theta)]$.

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.

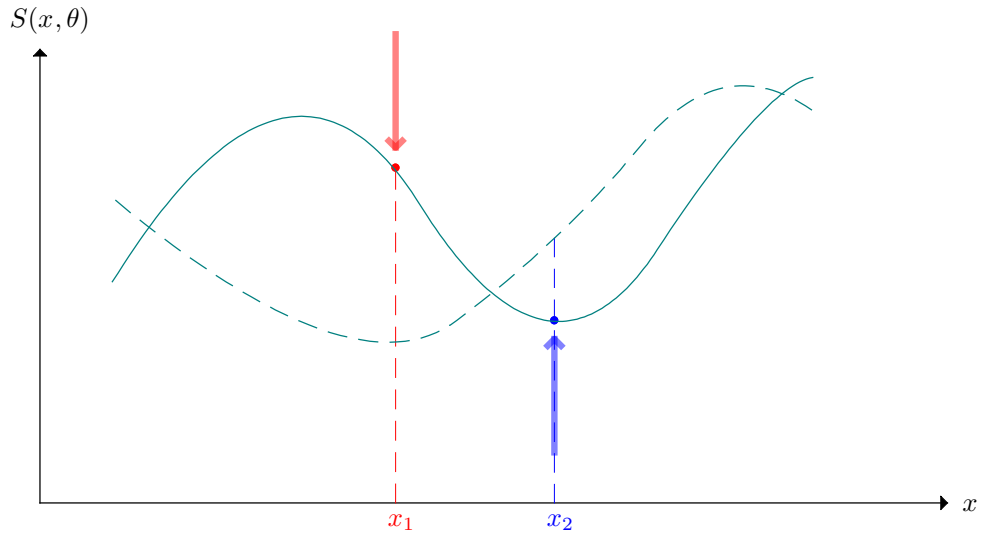

Figure 2. This figure illustrate how $\min_\theta L_{\mathrm{LA}}(\theta)$ will site a real world datum onto a local minimum of $S(\cdot,\theta)$. The green curve represents the current not-yet-optimized $S(\cdot,\theta)$. The x_1 (red point) is a real world datum while x_2 (blue point), which is currently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

Suppose that we have a set of raw data about an entity from classical physics. To describe the entity, 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

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

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 18. The action $S(x,\theta_\star)$ describes the dynamics extracted from the raw data.[11]

5.4. Example: Actions in Machine Learning (TODO)

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

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

For example, for a categorical classification task, $y\in\{1,\dots,M\}$ and $z\in\mathbb{R}^M$ for some M, and

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

And for regression task, $y,z\in\mathbb{R}^M$ for some M, and

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

Thus, we have an action

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

which is the loss per sample in machine learning.

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 $-\sum_{\alpha=1}^{M}(y^\alpha-f^\alpha(x;\theta))^2/2$.

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

In fact, equations (16), (17), and (19) can be regarded as an extremum of the functional

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

11. An experiment on general oscillators can be found in the oscillators/Oscillator.ipynb.

or explicitly

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

Indeed, variance on p gives equation (16).[12] Together with the partial derivative on μ, we get equation (17). Finally, partial derivative on θ directly gives equation (19).
Interestingly, the second term is just the $-L_{\mathrm{LA}}(\theta)$ in equation (20). So, the extremum is in fact a saddle point, as

$$(p_\star,\theta_\star,\mu_\star)=\min_{p,\mu}\max_{\theta}V(p,\theta,\mu). \tag{21}$$

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 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.

12. 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

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