1 Relative Entropy

1.1 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 $\{x_1, \ldots, x_n\}$ with p_i the probability of $X = x_i$. The Shannon entropy is thus a function of $p := (p_1, \ldots, p_n)$ defined as

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

where K is any positive constant. Interestingly, this expression is unique given some plausible conditions, which can be qualitatively expressed as [1]

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

1.2 Shannon Entropy Fails for Continuous Random Variable

The Shannon entropy, however, cannot be directly generalized to continuous random variable. Usually, entropy for continuous random variable X with distribution P is given by

$$H[p] := -K \int \mathrm{d}x\, p(x) \ln(p(x)),$$

which, however, is not well-defined, reflected by two issues. The first issue is that the p has dimension, indicated by $\int \mathrm{d}x\, p(x) = 1$. This means we put a dimensional quantity into logarithm, leading to a problem of dimension. The second issue is that the H defined so cannot be invariant under inverse map $X \to Y := \varphi(X)$ where φ is a diffeomorphism. As a "physical" quantity, H should be invariant under "non-physical" transformations, such as coordinates transformation characterized by the φ.

To eliminate the two issues, we shall extends the axiomatic description of entropy. The key to this extension is introducing another distribution, Q; and instead considering *the uncertainty (surprise) caused by P when prior knowledge is given by Q*. As we will see, this will solve the two issues altogether.

Explicitly, we extends the conditions as [2]

1. H is a continuous and local function of p and q;
2. H is invariant for diffeomorphic transformation; and
3. H can reduce to Shannon entropy when X is discrete and Q is uniform.

The first condition employs the locality of H, which is thought as natural since H has be a *functional*. The second condition is for solving the second issue. Finally, the third condition relates back to Shannon entropy.

Comparing with the conditions of Shannon entropy, the first condition is strengthened by adding locality; the second condition is absent since it is not well-defined for continuous variable.

1.3 Relative Entropy is Unique Solution to the Extended Conditions

Based on the first condition, H shall have the following expression

$$H[p, q] = \int \mathrm{d}x\, p(x)\, L(p(x), q(x)).$$

1. For details and quantitative description, see the appendix A of Jaynes (1957).
2. We follow the note by D. Rezende.

The second condition indicates that

$$L(p,q) = f(p/q)$$

for some continuous function f. Indeed, the $\mathrm{d}xp(x)$ is invariant under diffeomorphic transformation $X \to Y := \varphi(X)$ for some diffeomorphism φ. And if we take an infinitesimal transformation, we find $L(p(1+\epsilon), q(1+\epsilon)) = L(p,q)$, which indicates $\partial_1 L(p,q)\, p + \partial_2 L(p,q)\, q = 0$. Taking $p = Ae^t$ and $q = Be^t$, we find $L(Ae^t, Be^t) = C$ where C is a constant corresponding to t. This is valid only when $L(p,q) = f(p/q)$ for some f.

The third condition indicates that

$$f(x) \propto \ln x.$$

Indeed, when X is discrete and Q is uniform, we have $H[p,q] = \sum_i p_i f(p_i/q)$ which is compared with Shannon entropy $-K\sum_i p_i \ln(p_i)$, where $q = 1/n$ and n the alphabet size of X. For eliminating constraint $\sum_i p_i = 1$, we introduce ζ by $p_i = \mathrm{e}^{\zeta_i}/(\sum_j \mathrm{e}^{\zeta_j})$. Given j, taking derivative on ζ_j on both $\sum_i p_i f(p_i/q)$ and $-K\sum_i p_i \ln(p_i)$, we find

$$f(x_j) + x_j f'(x_j) - \frac{1}{n}\sum_{i=1}^{n} x_i[f(x_i) + x_i f'(x_i)] = -K\ln x_j + K\frac{1}{n}\sum_{i=1}^{n} x_i \ln x_i,$$

where $x_i := p_i/q$. Letting $g(x) := -x f(x)/K$, we arrive at

$$-g'(x_j) + \frac{1}{n}\sum_{i=1}^{n} x_i\, g'(x_i) = -\ln x_j + \frac{1}{n}\sum_{i=1}^{n} x_i \ln x_i.$$

TODO

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 [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 probability** $q_{t\to t'}(y|x)$ which describes the probability of transition from $X = x$ at time t to $X = y$ 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 probability at $X = x$ equals to the total probability that transits from any y with $y \neq x$ to x subtracting the total probability that transits from x to any y with $y \neq x$. That is, [4]

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

3. Such as the problem of Shannon entropy, which has no proper definition for continuous random variable.

4. Notice that in the case of $y = x$, the right hand side vanishes automatically. It is for this reason, the integral is over the whole alphabet $\mathcal{X}$.

which is called the **master equation**. [5] [6]

2.3 Detailed Balance Provides a Stationary Distribution

Let π a stationary solution of master equation 1. Then, π satisfies $\int_{\mathcal{X}} \mathrm{d}y\, [q_{\Delta t}(x|y)\, \pi(y) - q_{\Delta t}(y|x)\pi(x)] = 0$. But, this condition is too weak to be used. A more useful condition, which is stronger than this, is that the integrand vanishes everywhere. That is,

$$q_{\Delta t}(x|y)\, \pi(y) = q_{\Delta t}(y|x)\pi(x), \tag{3}$$

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

2.4 Detailed Balance with Ergodicity Monotonically Reduces Relative Entropy

Given the time t, if the time-dependent distribution $p(\cdot, t)$ and the stationary distribution π are both supported on $\mathcal{X}$, 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

$$H[p(\cdot,t), \pi] = \int_{\mathcal{X}} \mathrm{d}x p(x,t) \ln \frac{p(x,t)}{\pi(x)}. \tag{4}$$

It describes the uncertainty (surprise) caused by $p(\cdot, t)$ when prior knowledge is given by π. It is a plausible generalization of Shannon entropy to continuous random variables.

When $p(\cdot, t)$ is evolved by the master equation of $q_{\Delta t}$, to keep $H[p(\cdot,t), \pi]$ well-defined, we have to ensure that $p(\cdot, t)$ is supported on $\mathcal{X}$ for all t. This is guaranteed when $q_{\Delta t}(x|y) > 0$ for each x, $y \in \mathcal{X}$ and for each $\Delta t \in [0, T]$, where T is an arbitrary positive number. This property of transition probability is called **ergodicity**. Indeed, by repeatedly applying master equation 2, $p(x, t')$ is found to be supported on $\mathcal{X}$ for any $t' > t$. This keeps $H[p(\cdot,t), \pi]$ well-defined as long as it is well-defined initially.

---

5. There is another way of writing master equation, as

$$p(x, t + \Delta t) = \int_{\mathcal{X}} \mathrm{d}y q_{\Delta t}(x|y)\, p(y,t). \tag{2}$$

In fact, these two definitions are equivalent, which is the result of $\int_{\mathcal{X}} \mathrm{d}x p(x,t) = 1$. To make it transparent, we use the discrete version, as

$$p(x, t + \Delta t) = \sum_{y \in \mathcal{X}} q_{\Delta t}(x|y)\, p(y,t).$$

Since $\sum_{y \in \mathcal{X}} q_{\Delta t}(y|x) = 1$, we have $q_{\Delta t}(x|x) = 1 - \sum_{y \neq x} q_{\Delta t}(y|x)$. Thus,

$$\begin{aligned} p(x, t+\Delta t) - p(x,t) &= \sum_{y \in \mathcal{X}} q_{\Delta t}(x|y)\, p(y,t) - p(x,t) \\ &= \sum_{y \neq x} q_{\Delta t}(x|y)\, p(y,t) + q_{\Delta t}(x|x)\, p(x,t) - p(x,t) \\ \left\{ q_{\Delta t}(x|x) = 1 - \sum_{y \neq x} q_{\Delta t}(y|x) \right\} &= \sum_{y \neq x} [q_{\Delta t}(x|y)\, p(y,t) - q_{\Delta t}(y|x)\, p(x,t)] \\ &= \sum_{y \in \mathcal{X}} [q_{\Delta t}(x|y)\, p(y,t) - q_{\Delta t}(y|x)\, p(x,t)], \end{aligned}$$

which is the discrete version of master equation 1.

6. In many textures, master equation is defined by transition rate, instead of transition probability. This demands the smoothness of $q_{\Delta t}$ on Δt. But, this condition is not essential for applying master equation in many cases.

If $q_{\Delta t}$ is smooth on Δt, then by master equation 1, $p(\cdot,t)$ is smooth on t. Then we can calculate the time-derivative of relative entropy. 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 we can express the $\mathrm{d}H[p(\cdot,t),\pi]/\mathrm{d}t$ in a regular form, as [7]

$$H[p(\cdot,t+\mathrm{d}t),\pi]-H[p(\cdot,t),\pi]=-\frac{1}{2}\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y q_{\mathrm{d}t}(y|x)\,\pi(x)\left[\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right]\left[\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right]. \tag{5}$$

Since π is supported on $\mathcal{X}$, $q_{\mathrm{d}t}(y|x)\,\pi(x)$ cannot vanish everywhere. Thus, the sign of $\mathrm{d}H[p(\cdot,t),\pi]/\mathrm{d}t$ is determined by the last two terms. If $p(x,t)/\pi(x)>p(y,t)/\pi(y)$, then $\ln[p(x,t)/\pi(x)]>\ln[p(y,t)/\pi(y)]$, so that the whole expression is negative. The same 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; and this equation implies that $p(\cdot,t)=\pi$ since $\int_{\mathcal{X}}\mathrm{d}x\,p(x,t)=\int_{\mathcal{X}}\mathrm{d}x\,\pi(x)=1$. So, we conclude that

Theorem 1. *Suppose that the transition probability $q_{\Delta t}$ is ergodic and smooth on Δt. If there is a stationary distribution π supported on $\mathcal{X}$ such that detailed balance 3 holds, then for any time-dependent distribution $p(\cdot,t)$ initially supported on $\mathcal{X}$ and evolved by the master equation of $q_{\Delta t}$, $\mathrm{d}H[p(\cdot,t),\pi]/\mathrm{d}t$ is negative as long as $p(\cdot,t)\neq\pi$ and vanishes when $p(\cdot,t)=\pi$ for some t.*

This means the time-dependent distribution p will monotonically and constantly relax to the stationary distribution π.

Generally, we prove the monotonic reduction of relative entropy by using Fokker-Planck equation. With an integral by part, we arrive a negative definite expression, which means the monotonic reduction. This proof needs smooth structure on X, which is employed by 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).

---

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

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

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

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

Now, we replace $\partial p/\partial t$ by master equation in which Δt is replaced by the infinitesimal $\mathrm{d}t$, as

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

where $\mathrm{d}H[p,\pi]:=H[p(\cdot,t+\mathrm{d}t),\pi]-H[p(\cdot,t),\pi]$. Then, insert detailed balance $q_{\mathrm{d}t}(x|y)=q_{\mathrm{d}t}(y|x)\,\pi(x)/\pi(y)$, as

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

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

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

By adding the two previous results together, we find

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

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.

2.5 Temporal Smoothness of Transition Probability Is Necessary to Ensure Relaxation

The temporal smooth structure, however, cannot be avoided. Indeed, the smoothness of transition probability 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. [8]

To see this clearly, let us exam $H[p(\cdot,t+\Delta t),\pi]-H[p(\cdot,t),\pi]$ when Δt is not an infinitesimal. By definition,

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

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

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

The first line is recognized as $H[p(\cdot,t+\Delta t),p(\cdot,t)]$, which is non-negative. Following the same steps in section 2.4, 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 $p(\cdot,t)$ [9], 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 distribution P and of the transition rate W. This means, the conclusions in section 2 hold in this section, but the inverse is not true.

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.

8. 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, you can consider 1-dimensional case where the alphabet $\mathcal{X}=\{0,1\}$; the $p(x,t)$ is given by $\sigma(\zeta(t))$ where σ denotes the sigmoid function and $\zeta(t)$ is smooth on t. In this example, $p(x,t)$ is smooth on t but the random variable is discrete.

9. The difference is $\mathcal{O}(\Delta t^2)$.

Now, the master equation 1 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,

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

Altogether, we have

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

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

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

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

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

which is the k-order moment of $\epsilon\sim q_{\Delta t}(x+\epsilon|x)$. So, we arrive at

$$p(x,t+\Delta t)-p(x,t)=\sum_{k=1}^{+\infty}\frac{(-1)^k}{k!}\left(\frac{\partial}{\partial x^{\alpha_1}}\cdots\frac{\partial}{\partial x^{\alpha_k}}\right)[M^{\alpha^1\cdots\alpha^k}(x)\,p(x,t)],\tag{6}$$

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

Notice that deriving the Kramers–Moyal expansion needs the smoothness of $q_{\Delta t}(x|y)$ and $p(x,t)$ on x and y, but not the smoothness on Δt and t.

3.2 Langevin Dynamics that Satisfies Detailed Balance Is Conservative

Given a function $f:\mathbb{R}^n\to\mathbb{R}^n$ and a Wiener process $\mathrm{d}\xi$ with variance $2T\mathrm{d}t$, the **Langevin dynamics** is

$$\mathrm{d}X=f(X)\mathrm{d}t+\mathrm{d}\xi.$$

The transition probability $q_{\mathrm{d}t}(x+\epsilon|x)$ is $\mathcal{N}(f(x)\mathrm{d}t, 2T\mathrm{d}t)$ on ϵ, where T is a positive constant. Thus, moments $M^{\alpha}(x)=f^{\alpha}(x)\mathrm{d}t$, $M^{\alpha\beta}=2T\delta^{\alpha\beta}\mathrm{d}t$, and higher orders are of $o(\mathrm{d}t)$. The Kramers-Moyal expansion gives

$$\frac{\partial p}{\partial t}(x,t)=\nabla_{\alpha}(f^{\alpha}(x)\,p(x,t))+T\nabla^{2}p(x,t),$$

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 probability, we find that if there is a stationary distribution π such that Fokker-Planck equation satisfies detailed balance, then we must have [10]

$$f(x)=T\nabla\ln\pi(x). \tag{8}$$

This indicates that, to satisfy detailed balance, Langevin dynamics shall be conservative.

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. For this reason, we suggest that the objective to be maximized shall be the negative relative entropy instead of Shannon entropy. 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 $\int_{\mathcal{X}}\mathrm{d}x\,1<+\infty$, the negative relative entropy reduces to Shannon entropy. So, maximum-entropy principle shall minimize relative entropy.

Given a distribution Q of X that describes the prior knowledge, 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 $\{\mathbb{E}_p[f_\alpha]=\bar{f}_\alpha|\alpha=1,\ldots,m,\, f_\alpha\colon\mathcal{X}\to\mathbb{R}\}$. 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**. The P, thus, is the distribution which 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 $\int_{\mathcal{X}}\mathrm{d}x\,p(x)=1$. Also, 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

$$L[p,\lambda,\mu]:=\int_{\mathcal{X}}\mathrm{d}x\,p(x)\ln\frac{p(x)}{q(x)}+\lambda^{\alpha}\left(\int_{\mathcal{X}}\mathrm{d}x\,p(x)\,f_\alpha(x)-\bar{f}_\alpha\right)+\mu\left(\int_{\mathcal{X}}\mathrm{d}x\,p(x)-1\right). \tag{9}$$

10. To check detailed balance, we can employ either equation 3 or equation ?. Since $q_{\mathrm{d}t}$ is transparent, we check it using equation ?. That is, if 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 $\mathcal{N}(f(x)\mathrm{d}t, 2T\mathrm{d}t)$ on ϵ, the question becomes

$$\frac{1}{\sqrt{(4\pi T)^n}}\exp\left(-\frac{(\epsilon-f(x)\mathrm{d}t)^2}{4T\mathrm{d}t}\right)\pi(x)\overset{?}{=}\frac{1}{\sqrt{(4\pi T)^n}}\exp\left(-\frac{(-\epsilon-f(x-\epsilon)\mathrm{d}t)^2}{4T\mathrm{d}t}\right)\pi(x+\epsilon).$$

Expanding the both sides up to the first order of $\mathrm{d}t$ and ϵ, we directly find

$$f(x)=T\nabla\ln\pi(x). \tag{7}$$

So, we have

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

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

$$\frac{\partial \ln Z}{\partial \lambda^\alpha}(\lambda_\star) + \bar{f}_\alpha = 0 \tag{10}$$

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

$$Z(\lambda) := \int_{\mathcal{X}} \mathrm{d}x\, q(x) \exp(-\lambda^\alpha f_\alpha(x)); \tag{11}$$

and

$$p_\star(x) = p(x, \lambda_\star), \tag{12}$$

where

$$p(x,\lambda) := Z^{-1}(\lambda)\, q(x) \exp(-\lambda^\alpha f_\alpha(x)). \tag{13}$$

Notice that 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. It does not involve any parameter, which is the λ.

4.4 Weak Maximum-Entropy Principle Furnishes Observables

In many situations, we are not explicitly sure what the observables should be, but instead, we have some raw data, which can be described by an empirical distribution P_D. Remark that this empirical distribution is generally not supported on $\mathcal{X}$, since we may not have raw data for all possible values of X.

For example, in machine learning, we have raw data, but we cannot know what the observables should be on the raw data. In some cases, we have some ansatz for the observables, but in most cases, we have little. In fact, a major task of machine learning is to find out useful observables, which is also called **features**.

To solve this issue, we are to parameterize the observables, as $f(x,\varphi)\colon \mathcal{X} \times \mathbb{R}^p \to \mathbb{R}^m$, where $\mathbb{R}^p$ is parameter space. The observations are determined by the empirical distribution, as $\mathbb{E}_{p_D}[f(\cdot,\varphi)]$. And as a free parameter, we are to adjust φ to help minimize the relative entropy. So, the loss of maximum-entropy principle now becomes

$$L[p,\lambda,\varphi,\mu] = \int_{\mathcal{X}} \mathrm{d}x\, p(x) \ln\frac{p(x)}{q(x)} + \lambda^\alpha\Big(\int_{\mathcal{X}} \mathrm{d}x\, p(x) f_\alpha(x,\varphi) - \mathbb{E}_{p_D}[f_\alpha(\cdot,\varphi)]\Big) + \mu\Big(\int_{\mathcal{X}} \mathrm{d}x\, p(x) - 1\Big). \tag{14}$$

Following the same steps as before, denoting $\theta := (\lambda, \varphi)$ and

$$S(x,\theta) := \lambda^\alpha f_\alpha(x,\varphi),$$

the equation for an extremum $(p_\star, \theta_\star, \mu_\star)$ comes to be [11]

$$\frac{\partial \ln Z}{\partial \theta^\alpha}(\theta_\star) + \mathbb{E}_{p_D}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta_\star)\right] = 0, \tag{15}$$

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

$$Z(\theta) := \int_{\mathcal{X}} \mathrm{d}x\, q(x) \exp(-S(x, \theta)); \tag{16}$$

and we have

$$p_\star(x) = p(x, \theta_\star). \tag{17}$$

This can be seen as a weak version of maximum-entropy principle.

4.5 Extremum Can Be Found by Iteration Method

Generally, it is hard to solve the equation 10 and its weak version, equation 15. But, we can solve them numerically by iterative method. We are to show how to do this for the weak maximum-entropy principle, equation 15. Then, we will find that the equation 10 is solved also.

To solve equation 15 iteratively, we are to find a loss function for θ such that $\theta_\star$ locates at its minimum. Naturally, we consider the relative entropy $H[p_D, p(\cdot, \theta)]$ [12]. It is bounded below. Omitting the θ-independent terms, we get a loss function

$$L_{\text{iter}}(\theta) := \ln Z(\theta) + \mathbb{E}_{p_D}[S(\cdot, \theta)],$$

We find $\partial L_{\text{iter}} / \partial \theta^\alpha = 0$ gives equation 15. So, L_{iter} is a loss function of θ with $\theta_\star$ as a minimum; we can find the $\theta_\star$ by iteratively updating θ along the direction $-\partial L_{\text{iter}} / \partial \theta$.

With a series of direct calculus, we find

$$-\frac{\partial L_{\text{iter}}}{\partial \theta^\alpha}(\theta) = \mathbb{E}_{p(\cdot,\theta)}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right] - \mathbb{E}_{p_D}\left[\frac{\partial S}{\partial \theta^\alpha}(\cdot, \theta)\right]. \tag{18}$$

For equation 10, it becomes (now θ reduces to λ, and $\partial S / \partial \lambda^\alpha = f_\alpha$)

$$-\frac{\partial L_{\text{iter}}}{\partial \lambda^\alpha}(\lambda) = \mathbb{E}_{p(\cdot,\lambda)}[f_\alpha] - \bar{f}_\alpha. \tag{19}$$

4.6 When Is $\lambda_\star$ Solvable?

Even though it is hard to guarantee the equation 10 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 10, 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 simple calculation, we find

$$\frac{\partial^2 \ln Z}{\partial \lambda^\alpha \partial \lambda^\beta}(0) = \mathrm{Cov}_q(f_\alpha, f_\beta), \tag{20}$$

the covariance matrix of f under distribution q. TODO

---

11. Explicitly, we have

$$\begin{aligned}
\frac{\delta L}{\delta p(x)}[p, \lambda, \varphi, \mu] &= \ln p(x) + 1 - \ln q(x) + \lambda^\alpha f_\alpha(x, \varphi) + \mu \\
&= \ln p(x) + 1 - \ln q(x) + S(x, (\lambda, \varphi)) + \mu \\
\frac{\partial L}{\partial \lambda^\alpha}[p, \lambda, \varphi, \mu] &= \int_{\mathcal{X}} \mathrm{d}x\, p(x) f_\alpha(x, \varphi) - \mathbb{E}_{p_D}[f_\alpha(\cdot, \varphi)] \\
&= \int_{\mathcal{X}} \mathrm{d}x\, p(x) \frac{\partial S}{\partial \lambda^\alpha}(x, (\lambda, \varphi)) - \mathbb{E}_{p_D}\left[\frac{\partial S}{\partial \lambda^\alpha}(\cdot, (\lambda, \varphi))\right]; \\
\frac{\partial L}{\partial \varphi^\alpha}[p, \lambda, \varphi, \mu] &= \int_{\mathcal{X}} \mathrm{d}x\, p(x) \lambda^\beta \frac{\partial f_\beta}{\partial \varphi^\alpha}(x, \varphi) - \mathbb{E}_{p_D}\left[\lambda^\beta \frac{\partial f_\beta}{\partial \varphi^\alpha}(\cdot, \varphi)\right] \\
&= \int_{\mathcal{X}} \mathrm{d}x\, p(x) \frac{\partial S}{\partial \varphi^\alpha}(x, (\lambda, \varphi)) - \mathbb{E}_{p_D}\left[\frac{\partial S}{\partial \varphi^\alpha}(\cdot, (\lambda, \varphi))\right]; \\
\frac{\partial L}{\partial \mu}[p, \lambda, \varphi, \mu] &= \int_{\mathcal{X}} \mathrm{d}x\, p(x) - 1.
\end{aligned}$$

12. You may argue that p_D is not supported on $\mathcal{X}$, so that the relative entropy is not well-defined. But, we can consider the p_D as a limit of a smoothed empirical distribution, which is supported on $\mathcal{X}$.