Chapter 1

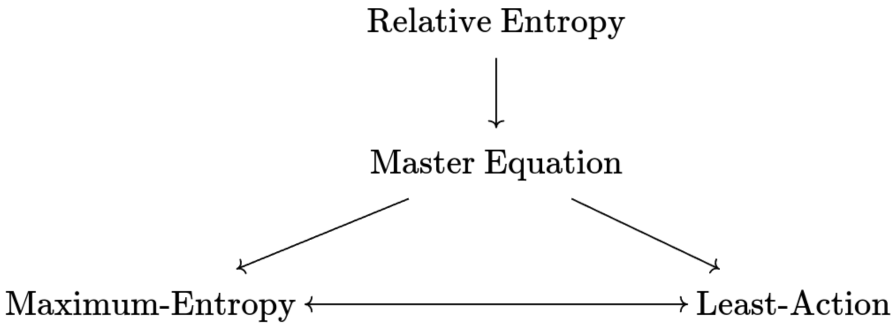

Figure 1.1. Scheme of sections. The solid arrow represents relation.

1.1 Relative Entropy

1.1.1 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]

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.

1.1.2 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.

[1.1] For details and quantitative description, see the appendix A of Jaynes (1957).

Explicitly, we extends the conditions as [1.2]

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.

1.1.3 Relative Entropy is Unique Solution to the Extended Conditions

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

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

The second condition indicates that

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

for some continuous function f. Indeed, the $\mathrm{d}x p(x)$ 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

$$f(x) \propto \ln x + C,$$

where C is a constant. 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 $H(p) := -K \sum_i p_i \ln(p_i)$, where $q = 1/n$ and n the alphabet size of X. We have $L(p) := H(p,q) + H(p) = \sum_i [p_i\, f(np_i) - K p_i \ln(p_i)]$. Take derivative on p_i, we find $\partial L / \partial p_i = 0$ implies $f(x) + x f'(x) - K \ln x + K(\ln n - 1) = 0$, where $x := np_i$. It has a unique solution with $f(1) = ?$

1.2 Master Equation, Detailed Balance, and Relative Entropy

1.2.1 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 [1.3] while the inverse will be safe and direct as long as any smooth structure of X is not employed throughout the discussion.

1.2.2 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 $q_{t \to t'}$.

[1.2] We follow the note by D. Rezende.

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

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, [1.4]

$$p(x,t+\Delta t)-p(x,t)=\int_{\mathcal{X}}\mathrm{d}y[q_{\Delta t}(x|y)p(y,t)-q_{\Delta t}(y|x)p(x,t)]. \tag{1.1}$$

which is called the **master equation**. [1.5] [1.6]

1.2.3 Detailed Balance Provides a Stationary Distribution

Let π a stationary solution of master equation 1.1. 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,

$$q_{\Delta t}(x|y)\,\pi(y)=q_{\Delta t}(y|x)\pi(x), \tag{1.3}$$

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

1.2.4 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

$$H[p(\cdot,t),\pi]=\int_{\mathcal{X}}\mathrm{d}x p(x,t)\ln\frac{p(x,t)}{\pi(x)}. \tag{1.4}$$

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. Based on equation 1.2, if $p(\cdot,t)$ has been supported on $\mathcal{X}$, $p(x,t+\Delta t)$ is also supported if and only if for each $x\in\mathcal{X}$, there is $y\in\mathcal{X}$ such that $q_{\Delta t}(x|y)>0$. That is, for each x, there is possibility that some y transits to x after a period Δt. This property of transition probability is called **ergodicity**. [1.7] By repeatedly applying master equation 1.2, $p(x,t')$ 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.

[1.4] 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 $\mathcal{X}$.

[1.5] There is another way of writing master equation, as

$$p(x,t+\Delta t)=\int_{\mathcal{X}}\mathrm{d}y q_{\Delta t}(x|y)\,p(y,t). \tag{1.2}$$

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 1.1.

[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.

[1.7] TODO: It seems that ergodicity is not defined as such.

If $q_{\Delta t}$ is smooth on Δt, then by master equation 1.1, $p(\cdot,t)$ 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 [1.8]

$$H[p(\cdot,t+\mathrm{d}t),\pi]-H[p(\cdot,t),\pi]=-\frac{1}{2}\int_{\mathcal{X}}\mathrm{d}x\int_{\mathcal{X}}\mathrm{d}y q_{\mathrm{d}t}(y|x)\,\pi(x)\left[\frac{p(x,t)}{\pi(x)}-\frac{p(y,t)}{\pi(y)}\right]\left[\ln\frac{p(x,t)}{\pi(x)}-\ln\frac{p(y,t)}{\pi(y)}\right]. \tag{1.5}$$

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 1.1. *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 1.3 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 1.3.2).

[1.8] 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.

1.2.5 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. [1.9]

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 1.2.4, 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 $p(\cdot,t)$ [1.10], 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.

1.3 Kramers-Moyal Expansion and Langevin Dynamics

We follow the discussion in section 1.2, 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 1.2 hold in this section, but the inverse is not true.

1.3.1 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 1.1 becomes

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

[1.9] 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.

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

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

$$p(x, t + \Delta t) - p(x, t) = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \left(\frac{\partial}{\partial x^{\alpha_1}} \cdots \frac{\partial}{\partial x^{\alpha_k}} \right) [M^{\alpha^1 \cdots \alpha^k}(x)\, p(x, t)], \tag{1.6}$$

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.

1.3.2 Langevin Dynamics that Satisfies Detailed Balance Is Conservative

Given $\mu: \mathbb{R}^n \to \mathbb{R}^n$ and $\Sigma: \mathbb{R}^n \to \mathbb{R}^{n \times n}$, which is positive definite and symmetric, the transition probability 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, moments $M^\alpha(x) = \mu^\alpha(x)\mathrm{d}t$, $M^{\alpha\beta}(x) = 2\Sigma^{\alpha\beta}(x)\mathrm{d}t$, and higher orders are of $o(\mathrm{d}t)$. The Kramers-Moyal expansion gives

$$\frac{\partial p}{\partial t}(x, t) = -\nabla_\alpha(\mu^\alpha(x)\, p(x, t)) + \nabla_\alpha \nabla_\beta(\Sigma^{\alpha\beta}(x)\, p(x, t)), \tag{1.7}$$

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 [1.11]

$$\mu^{\alpha}(x) = \Sigma^{\alpha\beta}(x)\nabla_{\beta}\left[\ln\pi(x) - \frac{1}{2}\operatorname{tr}\ln\Sigma(x)\right]. \tag{1.8}$$

This indicates that, to satisfy detailed balance, μ shall be conservative.[1.12]

1.4 Maximum-Entropy Principle

1.4.1 Conventions in This Section

Follow the conventions in section 1.2.

1.4.2 Maximum-Entropy Principle Shall Minimize Relative Entropy

As discussed in section 1.1, 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

$$L[p, \lambda, \mu] := \int_{\mathcal{X}}\mathrm{d}x\,p(x)\ln\frac{p(x)}{q(x)} + \lambda^{\alpha}\left(\int_{\mathcal{X}}\mathrm{d}x\,p(x)\,f_{\alpha}(x) - \bar{f}_{\alpha}\right) + \mu\left(\int_{\mathcal{X}}\mathrm{d}x\,p(x) - 1\right). \tag{1.9}$$

[1.11] 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 $\mathcal{O}(\sqrt{\Sigma(x)\,\mathrm{d}t})$, 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

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

[1.12] 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.

So, we have

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

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

$$\frac{\partial \ln Z}{\partial \lambda^\alpha}(\lambda_\star) + \bar{f}_\alpha = 0 \tag{1.10}$$

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

$$Z(\lambda) := \int_{\mathcal{X}} \mathrm{d}x\, q(x) \exp(-\lambda^\alpha f_\alpha(x)); \tag{1.11}$$

and

$$p_\star(x) = p(x, \lambda_\star), \tag{1.12}$$

where

$$p(x,\lambda) := Z^{-1}(\lambda)\, q(x) \exp(-\lambda^\alpha f_\alpha(x)). \tag{1.13}$$

Notice that the $\mu_\star$ has been included in the Z.

1.4.3 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.

1.4.4 When Is $\lambda_\star$ Solvable?

Even though it is hard to guarantee the equation 1.10 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 1.10, 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

$$\frac{\partial^2 \ln Z}{\partial \lambda^\alpha \partial \lambda^\beta}(0) = \mathrm{Cov}_q(f_\alpha, f_\beta), \tag{1.14}$$

the covariance matrix of f under distribution q. TODO

1.5 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.

1.5.1 Conventions in This Section

Follow the conventions in section 1.2.

1.5.2 Data Fitting Is Equivalent to Least-Action Principle

Let $p(\cdot,\theta)$ represent a parametrized distribution of X, and q a distribution of X that represents prior knowledge as in the case of maximum-entropy principle. Let $S(x,\theta):=-\ln(p(x,\theta)/q(x))-\ln Z(\theta)$ with Z to be determined. Notice that the distribution q is essential for defining S, since $\ln p(x,\theta)$ is not well-defined. [1.13] Then, we can re-formulate $p(x,\theta)$ as

$$p(x,\theta)=Z^{-1}(\theta)\,q(\theta)\exp(-S(x,\theta)), \tag{1.15}$$

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

$$Z(\theta)=\int_{\mathcal{X}}\mathrm{d}x\,q(\theta)\exp(-S(x,\theta)). \tag{1.16}$$

As a generic form of a parameterized distribution, it can be used to fit raw data that obeys an empirical distribution p_D, by adjusting parameter θ. To do so, we employ the usual loss function $H[p_D,p(\cdot,\theta)]$. By omitting the θ-independent terms, the loss function comes to be

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

The best fit $\theta_\star$ locates at the minimum of $L(\theta)$, where $p(\cdot,\theta_\star)=p_D$. At the minimum, we shall have $\partial L/\partial\theta=0$. Then, we find that $\theta_\star$ obeys

$$\mathbb{E}_{p(\cdot,\theta_\star)}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right]=\mathbb{E}_{p_D}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta_\star)\right]. \tag{1.17}$$

We can find the $\theta_\star$ by iteratively updating θ along the direction $-\partial L/\partial\theta$. With a series of direct calculus, we find

$$\frac{\partial L}{\partial\theta^\alpha}(\theta)=\mathbb{E}_{p_D}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta)\right]-\mathbb{E}_{p(\cdot,\theta)}\left[\frac{\partial S}{\partial\theta^\alpha}(\cdot,\theta)\right]. \tag{1.18}$$

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

$$L_{\mathrm{LA}}(\theta):=\mathbb{E}_{p_D}[S(\cdot,\theta)]-\mathbb{E}_{p(\cdot,\theta)}[S(\cdot,\theta)]. \tag{1.19}$$

The expectation $\mathbb{E}_{p(\cdot,\theta)}$ is computed by Monte-Carlo method. We sample data points from $p(\cdot,\theta)$ 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 1.2 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.

[1.13] First, when the random variable X is continuous, the $p(x,\theta)$ has dimension. But logarithm cannot accept a variable which has dimension. The second reason is that when we take a diffeomorphism of X to X', which can be viewed as a coordinate transformation, S cannot be invariant. But, to make it an action (later), S has to be invariant under coordinate transformation. For these two reasons, $\ln p(x,\theta)$ is not well-defined. But, it is easy to prove that $\ln(p(x,\theta)/q(x))$ is well-defined.

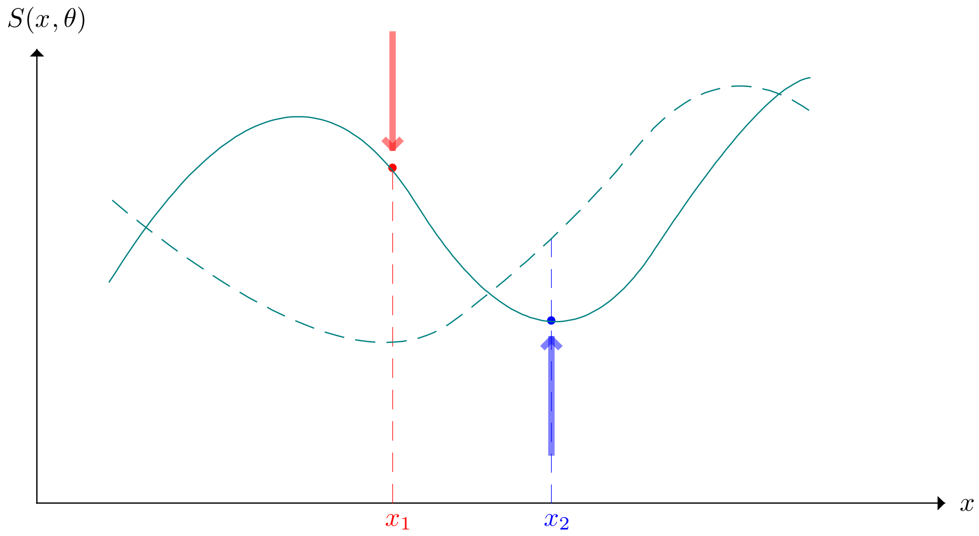

Figure 1.2. This figure illustrate how $\min_\theta L_{\text{LA}}(\theta)$ will site a real world datum onto a local minimum of $S(\cdot,\theta)$. The green curve represents the current not-yet-optimized $S(\cdot,\theta)$. The x_1 (red point) is a real world datum while x_2 (blue point), which is currently a local minimum of $S(\cdot,\theta)$, is not. Minimizing L_{LA} by tuning θ pushes the $\mathbb{E}_{p_D}[S(\cdot,\theta)]$ 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 $\mathbb{E}_{p(\cdot,\theta)}[S(\cdot,\theta)]$ 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.

1.5.3 Extract Dynamics from Raw Data: An Instance of Classical Physics

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 $\{x_k(1{:}T) | k=1,\ldots,D\}$ where time is discretized as $(1,\ldots,T)$ and the data size is D. We have employed $x_k(1{:}T)$ for the series of $(x_k(1),\ldots,x_k(T))$. Thus, each datum is a movie of the physical system, frame by frame. These raw data are obtained by experiments and measurements. Considering the errors in the measurements, variances shall be involved and the empirical distribution $p_D(x(1{:}T))$ comes to be a sum of normal distribution.

As a physical system, the q that represents free theory shall be Gaussian. It may be

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

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 [1.14]. Then, we find the best fit $\theta_\star$ by equation 1.18. The action $S[x,\theta_\star]$ describes the dynamics extracted from the raw data.[1.15]

[1.14] For example, $L(x,t,\theta) = f(x,\theta)$ where f is a neural network.

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

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

In fact, equations 1.15, 1.16, and 1.17 can be seen as an extremum of the functional

$$V[p,\theta,\mu]:=H[p,q]+(\mathbb{E}_p[S(x,\theta)]-\mathbb{E}_{p_D}[S(x,\theta)])+\mu(\mathbb{E}_p[1]-1),$$

or explicitly

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

Indeed, variance on p gives equation 1.15 and equation 1.16. And partial derivative on θ gives equation 1.17. Condition by partial derivative on μ has been involved in the $Z(\theta)$.

Interestingly, the second term is just the $-L_{\mathrm{LA}}(\theta)$ in equation 1.19. So, the extremum is in fact a saddle point, as

$$(p_\star,\theta_\star,\mu_\star)=\min_{p,\mu}\max_{\theta}V[p,\theta,\mu]. \tag{1.20}$$

The minimization minimizes the relative entropy between p and q and the expectation of action $S(\cdot,\theta)$ by tuning p, which in turn relates the probability p with the action $S(\cdot,\theta)$. The maximization sites real data onto the action's local minima by tuning θ. So, we find that maximum-entropy principle and least-action principle are saddle point of a functional V.

1.5.5 Actions in Machine Learning

As figure 1.2 indicates, we shall push down the real world data while pull up the data sampled from the $p(x,\theta)$, until the two forces balanced. In fact, to sample from $p(x,\theta)$, we will not fully evaluate the Markov chain Monte-Carlo to equilibrium, which will consume a plenty of computation resources, but only run several steps. In this case, the data sampled from $p(x,\theta)$ will be close to the initial of the Markov chain Monte-Carlo, for which we employ the real world data. So, let $x\sim p_D$ as the real word datum, we have the sampled $x'\approx x$. The difference $\Delta x:=x'-x$ is small enough, so that we have the approximation of the equivalent loss L_{LA} (equation 1.19) as

$$\begin{aligned}L_{\mathrm{LA}}(\theta)&=\mathbb{E}_{p_D}[S(\cdot,\theta)]-\mathbb{E}_{p(\cdot,\theta)}[S(\cdot,\theta)]\\&\approx\mathbb{E}_{x\sim p_D}[S(x,\theta)-S(x+\Delta x,\theta)]\\&=\mathbb{E}_{x\sim p_D}\left[-\frac{\partial S}{\partial x^\alpha}(x,\theta)\Delta x^\alpha\right].\end{aligned}$$

If we use Langevin dynamics for Markov chain Monte-Carlo with extremely low temperature (section 1.3.2), and run for a single-step, we will have $\Delta x^\alpha\approx-(\partial S/\partial x_\alpha)(x,\theta)\,\Delta t$, where Δt is the step-size. Plugging back to L_{LA}, we have

$$L_{\mathrm{LA}}(\theta)\approx\tilde{L}_{\mathrm{LA}}(\theta):=\mathbb{E}_{x\sim p_D}\left[\left\|\frac{\partial S}{\partial x}\right\|_2^2(x,\theta)\right]\Delta t.$$

Minimizing $\tilde{L}_{\mathrm{LA}}(\theta)$ by adjusting θ is approximately reducing the norm of $\partial S/\partial x$ on real word data. This method of optimization is quite different from that used in machine learning. In machine learning, the action turns to be the loss function that characterizes the difference between the targets and the model predictions. The aim of machine learning is minimizing the action (loss function) instead of the norm of its gradient.

There are two kinds of tasks in machine learning: regression and classification. For regression task, the loss function that is usually employed is mean squared error. And for classification, the loss function is chosen to be cross-entropy. Let $f(x,\theta)$ the model with parameter θ and the input-target pair $(x,y)\sim p_D$, we are to compute

$$\frac{\tilde{L}_{\mathrm{LA}}(\theta)}{\Delta t}=\mathbb{E}_{(x,y)\sim p_D}\left[\left\|\frac{\partial S}{\partial x}\right\|_2^2(x,y,\theta)+\left\|\frac{\partial S}{\partial y}\right\|_2^2(x,y,\theta)\right]$$

for these two loss functions.

In regression task, we have model input $x\in\mathbb{R}^n$ and scalar target $y\in\mathbb{R}$, where $n\geqslant 1$. Mean squared error is defined by $\mathbb{E}_{(x,y)\sim p_D}[(f(x,\theta)-y)^2]$. So, the action is

$$S_{\mathrm{MSE}}(x,y,\theta):=(f(x,\theta)-y)^2.$$

Directly, we have

$$\frac{\partial S_{\mathrm{MSE}}}{\partial x^{\alpha}}(x, y, \theta) = 2(f(x, \theta) - y)\frac{\partial f}{\partial x^{\alpha}}(x, \theta);$$
$$\frac{\partial S_{\mathrm{MSE}}}{\partial y}(x, y, \theta) = 2(y - f(x, \theta)).$$

Thus, for mean squared error,

$$\frac{\tilde{L}_{\mathrm{LA}}(\theta)}{\Delta t} = 4\mathbb{E}_{(x,y)\sim p_D}\left[S_{\mathrm{MSE}}(x, y, \theta)\left(1 + \left\|\frac{\partial f}{\partial x}\right\|_2^2(x, \theta)\right)\right].$$

So, minimizing $\tilde{L}_{\mathrm{LA}}(\theta)$ by adjusting θ will minimize S_{MSE} itself as well as the norm of $\partial f / \partial x$ on real world data p_D. The norm of $\partial f / \partial x$ can be viewed as a regularization term, which provides a greater robustness for the model.

In classification task, we have model input $x \in \mathbb{R}^n$ and categorical probabilistic target $y \in \mathbb{R}^m$, where $n \geqslant 1$ and $m > 1$. Cross-entropy is defined as $\mathbb{E}_{(x,y)\sim p_D}[-\sum_{\alpha} y_{\alpha} \ln q_{\alpha}(x, \theta)]$, where $q_{\alpha}(x, \theta) := \exp(f^{\alpha}(x, \theta)) / \sum_{\beta} \exp(f^{\beta}(x, \theta))$. So, the action is

$$S_{\mathrm{CE}}(x, y, \theta) := -\sum_{\alpha} y^{\alpha} \ln \frac{\exp(f^{\alpha}(x, \theta))}{\sum_{\beta} \exp(f^{\beta}(x, \theta))}.$$

Directly, we have

$$\frac{\partial S_{\mathrm{CE}}}{\partial x^{\alpha}}(x, y, \theta) = \sum_{\beta} (q_{\beta} - y_{\beta})\frac{\partial f^{\beta}}{\partial x^{\alpha}};$$
$$\frac{\partial S_{\mathrm{CE}}}{\partial y^{\alpha}}(x, y, \theta) = -\ln q_{\alpha}.$$

Thus, for cross-entropy,

$$\frac{\tilde{L}_{\mathrm{LA}}(\theta)}{\Delta t} = \mathbb{E}_{(x,y)\sim p_D}\left[\left\|\sum_{\alpha} (y_{\alpha} - q_{\alpha})\frac{\partial f^{\alpha}}{\partial x}\right\|_2^2 + \|\ln q\|_2^2\right].$$

Minimizing $\tilde{L}_{\mathrm{LA}}(\theta)$ by adjusting θ will minimize both of the terms in the expectation. On the extremum $\theta_{\star}$ where $\tilde{L}_{\mathrm{LA}}(\theta_{\star}) = 0$, these terms shall vanish. But, the second term will never vanish. This implies that, we cannot expect that the real world data locate in the bottoms of loss function.[1.16]

As a summary, for regression task, the real world data locate in the bottoms of loss function of the best fit model, but never for classification task. Even though, we can employ mean squared error for classification, as for regression.[1.17]

[1.16]. You may challenge that we shall define y as logits instead of probabilities, thus

$$S_{\mathrm{CE}}(x, y, \theta) = -\sum_{\alpha} \frac{\exp(y^{\alpha})}{\sum_{\beta} \exp(y^{\beta})} \ln \frac{\exp(f^{\alpha}(x, \theta))}{\sum_{\beta'} \exp(f^{\beta'}(x, \theta))}.$$

Let $p^{\alpha}(y) := \exp(y^{\alpha}) / \sum_{\beta} \exp(y^{\beta})$, we have

$$\frac{\partial S_{\mathrm{CE}}}{\partial x^{\alpha}}(x, y, \theta) = \sum_{\beta} (q_{\beta} - p_{\beta})\frac{\partial f^{\beta}}{\partial x^{\alpha}};$$
$$\frac{\partial S_{\mathrm{CE}}}{\partial y^{\alpha}}(x, y, \theta) = p_{\alpha}\left(\sum_{\beta} p_{\beta} \ln q_{\beta} - \ln q_{\alpha}\right).$$

Thus, for cross-entropy,

$$\frac{\tilde{L}_{\mathrm{LA}}(\theta)}{\Delta t} = \mathbb{E}_{(x,y)\sim p_D}\left[\left\|\sum_{\alpha} (p_{\alpha} - q_{\alpha})\frac{\partial f^{\alpha}}{\partial x}\right\|_2^2 + \left\|p\left(\ln q - \sum_{\alpha} p_{\alpha} \ln q_{\alpha}\right)\right\|_2^2\right].$$

Minimizing $\tilde{L}_{\mathrm{LA}}(\theta)$ by adjusting θ will minimize both of the terms in the expectation. On the extremum $\theta_{\star}$ where $\tilde{L}_{\mathrm{LA}}(\theta_{\star}) = 0$, these terms shall vanish. Since p cannot be one-hot, when the second term vanishes, $\ln q_{\beta} = \sum_{\alpha} p_{\alpha} \ln q_{\alpha}$ for all β. It means that the components of $\ln q$ are all equal. By the definition of q, it in turn means that the components of $f(x, \theta_{\star})$ are all equal, for each x sampled from p_D. Then, to make the first term vanish, we have $(\partial f / \partial x)(x, \theta_{\star}) = 0$ for each x sampled from p_D, unless the components of p are all equal, which is impossible. Apparently, the extremum of $\tilde{L}_{\mathrm{LA}}(\theta)$ is not the extremum of the loss in machine learning, $\mathbb{E}_{(x,y)\sim p_D}[S_{\mathrm{CE}}(x, y, \theta)]$. In other words, for the best fit classification model that minimizes the loss, we cannot expect that the real world data locate in the bottoms of loss function. We now arrive at the same conclusion as before.

[1.17]. Experiments can be found in the folder actions.