

1 Lyapunov Function

Notation 1. Overall notations in this section are:

- \mathcal{M} a manifold, and μ its measure, e.g. $\mu(x) = \sqrt{g(x)}$ if \mathcal{M} is Riemannian with metric g_{ab} ;
- if $p(x)$ the distribution of random variable X , then

$$\langle f \rangle_p = \langle f \rangle_X := \int_{\mathcal{M}} d\mu(x) p(x) f(x);$$

- if D is a set of samples, then

$$\langle f \rangle_D := \frac{1}{|D|} \sum_{x \in D} f(x);$$

- let $\mathcal{N}(\mu, \Sigma)$ denotes normal distribution with mean μ and covariance Σ ;
- given function g , let $f\{g\}$, or $f_{\{g\}}$, denote a function constructed out of g , that is,

$$f\{\cdot\}: (\mathcal{M} \rightarrow A) \rightarrow (\mathcal{M} \rightarrow B).$$

1.1 Relaxation

Next, we illustrate how, during a non-equilibrium process, a distribution p relaxes to its stationary distribution q , and how this process relates to the variational inference. Further, we try to find the most generic dynamics that underlies the non-equilibrium to equilibrium process, on both macroscopic (distribution) and microscopic (“particle”) viewpoints.

First, we shall define what relaxation is, via free energy.

Definition 2. [Free Energy]

Let $E(x): \mathcal{M} \rightarrow \mathbb{R}$. Define stationary distribution

$$q_E(x) := \frac{\exp(-E(x)/T)}{Z},$$

where $T > 0$ and $Z := \int_{\mathcal{M}} d\mu(x) \exp(-E(x)/T)$. Given E , for any time-dependent distribution $p(x, t)$, define free energy as

$$F_E[p(\cdot, t)] := T D_{\text{KL}}(p \| q_E) - T \ln Z = T \int_{\mathcal{M}} d\mu(x) p(x, t) \ln \frac{p(x, t)}{q_E(x)} - T \ln Z.$$

Or, equivalently,

$$F_E[p(\cdot, t)] := \langle E \rangle_{p(\cdot, t)} - TH[p(\cdot, t)],$$

where entropy functional $H[p(\cdot, t)] := \langle -\ln p(\cdot, t) \rangle_p$.

Definition 3. [Relaxation]

For a time-dependent distribution $p(x, t)$ on \mathcal{M} , we say p relaxes to q_E if and only if the free energy $F_E[p(\cdot, t)]$ monotonically decreases to its minimum, where $p(\cdot, t) = q_E$.

We can visualize this relaxation process by an imaginary ensemble of juggling “particles” (or “bees”). Initially, they are arbitrarily positioned. This forms a distribution of “particles” p . With some underlying dynamics, these “particles” moves and finally the distribution relaxes, if it can, to a stationary distribution q_E . Apparently, the underlying dynamics and the E are correlated. We first provide a way of peeping the underlying dynamics, that is, the “flux”.

Lemma 4. [Conservation of “Mass”]

For any time-dependent distribution $p(x, t)$, there exists a “flux” $f^a\{p\}(x, t)$ s.t.

$$\frac{\partial p}{\partial t}(x, t) + \nabla_a(f^a\{p\}(x, t) p(x, t)) = 0.$$

What is the dynamics of p by which any initial p will finally relax to q_E ? That is, what is the sufficient (and essential) condition of relaxing to q_E for any p ? Because of the conservation of “mass”, the dynamics of p , i.e. $\partial p / \partial t$, is determined by a “flux”, f^a . Thus, this sufficient (and essential) condition must be about the f^a .

Lemma 5. *Given p and (x, t) , for any $f^a\{p\}(x, t)$, we can always construct a $K^{ab}\{p\}(x, t)$ s.t.*

$$f^a\{p\}(x, t) = -K^{ab}\{p\}(x, t) \nabla_b \{T \ln p(x, t) + E(x)\}.$$

Proof. For any vector f^a and v_a , we can always construct a tensor K^{ab} s.t. $f^a = K^{ab} v_b$. Indeed, we can rotate v_a to the direction of f^a and then dimension-wise rescale to f^a . This rotation and dimension-wise rescaling compose the linear transform K^{ab} . Now, letting

$$v_a = -\nabla_a \{T \ln p(x, t) + E(x)\},$$

we arrive at the conclusion. \square

Now, we claim a sufficient condition of relaxing to q_E for any p .

Theorem 6. *[Fokker-Planck Equation]*

If, for any p and t , the symmetric part of $K^{ab}\{p\}(x, t)$ is almost everywhere positive definite on \mathcal{M} , then any p evolves by this “flux” will relax to q_E .

Proof. Directly

$$\begin{aligned} \frac{dF_E}{dt}[p(\cdot, t)] &= T \int_{\mathcal{M}} d\mu(x) \frac{\partial p}{\partial t}(x, t) \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \\ \{\text{Conservation of mass}\} &= -T \int_{\mathcal{M}} d\mu(x) \nabla_a [f^a\{p\}(x, t) p(x, t)] \left[\ln \frac{p(x, t)}{q(x)} + 1 \right]. \end{aligned}$$

Since

$$\nabla_a [f^a\{p\}(x, t) p(x, t)] \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] = \nabla_a \left\{ [f^a\{p\}(x, t) p(x, t)] \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \right\} - [f^a\{p\}(x, t) p(x, t)] \nabla_a \left[\ln \frac{p(x, t)}{q(x)} + 1 \right],$$

we have

$$\begin{aligned} \frac{dF_E}{dt}[p(\cdot, t)] &= -T \int_{\mathcal{M}} d\mu(x) \nabla_a [f^a\{p\}(x, t) p(x, t)] \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \\ &= -T \int_{\mathcal{M}} d\mu(x) \nabla_a \left\{ [f^a\{p\}(x, t) p(x, t)] \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \right\} \\ &\quad + T \int_{\mathcal{M}} d\mu(x) [f^a\{p\}(x, t) p(x, t)] \nabla_a \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \\ [\text{Divergence theorem}] &= -T \int_{\partial\mathcal{M}} dS_a p(x, t) f^a\{p\}(x, t) \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \\ &\quad + T \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) \nabla_a \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \end{aligned}$$

The first term vanishes.¹ Then, direct calculus shows

$$\begin{aligned} \frac{dF_E}{dt}[p(\cdot, t)] &= T \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) \nabla_a \left[\ln \frac{p(x, t)}{q(x)} + 1 \right] \\ &= T \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) [\nabla_a \ln p(x, t) - \nabla_a \ln q(x)] \\ \{q(x) := \dots\} &= \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) [T \nabla_a \ln p(x, t) + \nabla_a E(x)] \\ &= \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) \nabla_a \{T \ln p(x, t) + E(x)\}. \end{aligned}$$

By the previous lemma, we have

$$\begin{aligned} \frac{dF_E}{dt}[p(\cdot, t)] &= \int_{\mathcal{M}} d\mu(x) p(x, t) f^a\{p\}(x, t) \nabla_a \{T \ln p(x, t) + E(x)\} \\ \{f^a = \dots\} &= - \int_{\mathcal{M}} d\mu(x) p(x, t) K^{ab}\{p\}(x, t) \nabla_a \{T \ln p(x, t) + E(x)\} \nabla_b \{T \ln p(x, t) + E(x)\}. \end{aligned}$$

Letting $S^{ab} := (K^{ab} + K^{ba})/2$ and $A^{ab} := (K^{ab} - K^{ba})/2$, we have $K^{ab} = S^{ab} + A^{ab}$, where S^{ab} is symmetric and A^{ab} anti-symmetric. Then,

$$\begin{aligned} \frac{dF_E}{dt}[p(\cdot, t)] &= - \int_{\mathcal{M}} d\mu(x) p(x, t) [S^{ab}\{p\}(x, t) + A^{ab}\{p\}(x, t)] \nabla_a \{T \ln p(x, t) + E(x)\} \nabla_b \{T \ln p(x, t) + E(x)\} \\ \{A^{ab} = A^{ba}\} &= - \int_{\mathcal{M}} d\mu(x) p(x, t) S^{ab}\{p\}(x, t) \nabla_a \{T \ln p(x, t) + E(x)\} \nabla_b \{T \ln p(x, t) + E(x)\}. \end{aligned}$$

The condition claims that $S^{ab}\{p\}(x, t)$ is positive definite for any p and (x, t) . Then, the integrand is a positive definite quadratic form, being positive if and only if $\nabla_a \{T \ln p(x, t) + E(x)\} \neq 0$. Then, we find $(dF_E/dt)[p(\cdot, t)] < 0$ as long as $\nabla_a \{T \ln p(x, t) + E(x)\} \neq 0$ at some x , i.e. $p \neq q$, and $(dF_E/dt)[p(\cdot, t)] = 0$ if and only if $\nabla_a \{T \ln p(x, t) + E(x)\} = 0$ for $\forall x$, i.e. $p = q$. Thus proof ends. \square

¹ **To-do: Explain the reason explicitly.**

Remark 7. [Sufficient but Not Essential]

However, this is not an essential condition of relaxing to q_E for any p . Indeed, we proved the integrand of $(dF_E/dt)[p(\cdot, t)]$ is negative everywhere, which implies the integral, i.e. $(dF_E/dt)[p(\cdot, t)]$, is negative. But, we cannot exclude the case where the integrand is not negative everywhere, whereas the integral is still negative. During the proof, this is the only place that leads to the non-essential-ness, which is hard to overcome.

As the dynamics of distribution is a macroscopic viewpoint, the microscopic viewpoint, i.e. the stochastic dynamics of single “particle”², is as follow.

Theorem 8. [Stochastic Dynamics]

If K^{ab} is symmetric, independent of p and almost everywhere smooth on \mathcal{M}^3 , then Fokker-Planck equation is equivalent to the stochastic dynamics

$$dx^a = [T \nabla_b K^{ab}(x, t) - K^{ab}(x, t) \nabla_b E(x)] dt + \sqrt{2T} dW^a(x, t),$$

where

$$dW \sim \mathcal{N}(0, K(x, t) dt).$$

Proof. From the difference of the stochastic dynamics,

$$\Delta x^a = [T \nabla_b K^{ab}(x, t) - K^{ab}(x, t) \nabla_b E(x)] \Delta t + \sqrt{2T} \Delta W^a(x, t),$$

by Kramers–Moyal expansion ¹⁷, we have

$$p(x, t + \Delta t) - p(x, t) = \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!} \nabla_{a_1} \cdots \nabla_{a_n} [p(x, t) \langle \Delta x^{a_1} \cdots \Delta x^{a_n} \rangle_{\Delta x}].$$

For $n=1$, since $\langle dW^a(x, t) \rangle_{dW} = 0$, the term is $-\nabla_a [p(x, t) \langle \Delta x^a \rangle_{\Delta x}] = \nabla_a \{p(x, t) [K^{ab}(x, t) \nabla_b E(x) - T \nabla_b K^{ab}(x, t)]\} \Delta t$. And for $n=2$, up to $o(\Delta t)$, only $T \nabla_a \nabla_b [p(x, t) K^{ab}(x, t)] \Delta t$ left. For $n \geq 3$, all are $o(\Delta t)$. So, we have

$$\frac{p(x, t + \Delta t) - p(x, t)}{\Delta t} = \nabla_a \{p(x, t) [K^{ab}(x, t) \nabla_b E(x) - T \nabla_b K^{ab}(x, t)]\} + T \nabla_a \nabla_b \{p(x, t) K^{ab}(x, t)\} + o(1).$$

Letting $\Delta t \rightarrow 0$, we find

$$\begin{aligned} \frac{\partial p}{\partial t}(x, t) &= \nabla_a \{p(x, t) [K^{ab}(x, t) \nabla_b E(x) - T \nabla_b K^{ab}(x, t)]\} + T \nabla_a \nabla_b \{p(x, t) K^{ab}(x, t)\} \\ \{\text{Expand}\} &= \nabla_a \{K^{ab}(x, t) \nabla_b E(x) p(x, t)\} - \nabla_a \{T \nabla_b K^{ab}(x, t) p(x, t)\} \\ &\quad + \nabla_a \{T K^{ab}(x, t) \nabla_b p(x, t)\} + \nabla_a \{T \nabla_b K^{ab}(x, t) p(x, t)\} \\ &= \nabla_a \{K^{ab}(x, t) \nabla_b E(x) p(x, t)\} + \nabla_a \{T K^{ab}(x, t) \nabla_b p(x, t)\}, \end{aligned}$$

which is just the Fokker-Planck equation. Indeed, the Fokker-Planck equation ⁶ is

$$\begin{aligned} \frac{\partial p}{\partial t}(x, t) &= -\nabla_a (f^a \{p\}(x, t) p(x, t)) \\ \{f^a = \dots\} &= \nabla_a (K^{ab} \{p\}(x, t) \nabla_b \{T \ln p(x, t) + E(x)\} p(x, t)) \\ \{K^{ab} \text{ independent of } p\} &= \nabla_a (K^{ab}(x, t) \nabla_b \{T \ln p(x, t) + E(x)\} p(x, t)) \\ \{\text{Expand}\} &= \nabla_a \{K^{ab}(x, t) \nabla_b E(x) p(x, t)\} + \nabla_a \{T K^{ab}(x, t) \nabla_b p(x, t)\}, \end{aligned}$$

exactly the same. Thus proof ends. \square

Question 1. Given a stochastic dynamics, how can we determine if there exists the E , or the stationary distribution q_E ?

Question 2. Further, if it exists, then how can we reveal it? Precisely, in the case $T \rightarrow 0$, given $(dx^a/dt) = h^a(x, t)$, how can we reconstruct the E and find a positive definite K^{ab} , s.t. $h^a(x) = K^{ab}(x, t) \nabla_b E(x)$?

1.2 Minimize Free Energy Principle

In the real world, there can be two types of variables: ambient and latent. The ambient variables are those observed directly, like sensory inputs or experimental observations. While the latent are usually more simple and basic aspects, like wave-function in QM.

We formulate the E as a function of $(v, h) \in \mathcal{V} \times \mathcal{H}$, where v , for visible, represents the ambient and h , for hidden, represents the latent. Then we have

Lemma 9. [Conditional Free Energy]

². For the conception of stochastic dynamics, c.f. B.2.

³. **TODO: Check this.**

Given v , if define

$$Z(v) := \int_{\mathcal{H}} dh \exp(-E(v, h)/T),$$

then we have a (conditional) free energy of distribution $p(h)$

$$\begin{aligned} F_E[p|v] &:= TD_{\text{KL}}(p||q_E(\cdot|v)) - T \ln Z(v) \\ &= \langle E(v, \cdot) \rangle_p - TH[p]. \end{aligned}$$

Ansatz 10. [Minimize Free Energy Principle]

Let $p(h)$ the latent distribution. On one hand, we want to locate it to the minimum of E . That is, given the ambient v , we want to minimize $\langle E(v, \cdot) \rangle_p$, where we have marginalized the latent. On the other hand, we shall keep the minimal prior knowledge on the latent, that is, maximize $H[p]$. So, we minimize $\langle E(v, \cdot) \rangle_p - TH[p]$, where the positive constant T balances the two aspects. This happens to be the (conditional) free energy.

Lemma 11. If E is in a function family parameterized by $\theta \in \mathbb{R}^N$, then we have

$$\frac{\partial}{\partial \theta^\alpha} \{-T \ln Z(v)\} = \left\langle \frac{\partial E}{\partial \theta^\alpha}(v, \cdot) \right\rangle_{q_E(\cdot|v)}.$$

Thus, we propose an EM-like algorithm that minimizes the free energy, as

Algorithm 12. [Recall and Learn (RL)]

To minimize free energy $F_E[p|v]$, we have two steps:

1. minimize $\langle E(v, \cdot) \rangle_p - TH[p]$ by the stochastic dynamics until relaxation, where $p = q_E(\cdot|v)$; then
2. minimize $-T \ln Z(v)$ by gradient descent and replacing $\langle (\partial E / \partial \theta^\alpha)(v, \cdot) \rangle_{q_E(\cdot|v)} \rightarrow \langle (\partial E / \partial \theta^\alpha)(v, \cdot) \rangle_p$.

By repeating these two steps, we get smaller and smaller free energy.

For instance, in a brain, the first step can be illustrated as recalling, and the second as learning (searching for a more proper memory, or code of information).

1.3 Example: Continuous Hopfield Network

Here, we provide a biological inspired example, for illustrating both the stochastic dynamics 8 and the RL algorithm 12.

Definition 13. [Continuous Hopfield Network]⁴

Let $U^{\alpha\beta}$ and I^α constants, and $L_v(v)$ and $L_h(h)$ scalar functions. Define $f_\alpha := \partial_\alpha L_h$, $g_\alpha := \partial_\alpha L_v$. Then the dynamics of continuous Hopfield network is defined as

$$\begin{aligned} \frac{dv^\alpha}{dt} &= U^{\alpha\beta} f_\beta(h) - v^\alpha + I^\alpha; \\ \frac{dh^\alpha}{dt} &= U^{\beta\alpha} g_\beta(v) - h^\alpha, \end{aligned}$$

where U describes the strength of connection between neurons, and f, g the activation functions of latent and ambient, respectively. Further, we have the E constructed as

$$E(v, h) = [(v^\alpha - I^\alpha) g_\alpha(v) - L_v(v)] + [h^\alpha f_\alpha(h) - L_h(h)] - U_{\alpha\beta} g^\alpha(v) f^\beta(h).$$

Next, we convert this deterministic dynamics to its stochastic version.

Theorem 14. If $f = \partial L_h$ and $g = \partial L_v$ are piecewise linear functions⁵, and the Hessian matrix of L_v and L_h are positive definite, then the stochastic dynamics of the continuous Hopfield network is

$$\begin{aligned} \frac{dv^\alpha}{dt} &= K_v^{\alpha\beta}(v) [U_{\beta\gamma} f^\gamma(h) - v_\beta + I_\beta] + \sqrt{2T} dW_v^\alpha; \\ \frac{dh^\alpha}{dt} &= K_h^{\alpha\beta}(h) [U_{\gamma\beta} g^\gamma(v) - h^\beta] + \sqrt{2T} dW_h^\alpha, \end{aligned}$$

4. Originally illustrated in Large Associative Memory Problem in Neurobiology and Machine Learning, Dmitry Krotov and John Hopfield, 2020.

5. E.g. LeakyReLU.

where $K_v(v) := [\partial^2 L_v(v)]^{-1}$ and $K_h(h) := [\partial^2 L_h(h)]^{-1}$ are piecewise constant matrices.⁶

Proof. Directly, we have

$$\begin{aligned} \frac{\partial E}{\partial v^\alpha}(v, h) &= g_\alpha(v) + (v^\beta - I^\beta) \frac{\partial g_\beta}{\partial v^\alpha}(v) - \frac{\partial L_v}{\partial v^\alpha}(v) - U^{\beta\gamma} f_\gamma(h) \frac{\partial g_\beta}{\partial v^\alpha}(v) \\ \left\{ g_\alpha = \frac{\partial L_v}{\partial v^\alpha} \right\} &= -[U^{\beta\gamma} f_\gamma(h) + v^\beta - I^\beta] \frac{\partial g_\beta}{\partial v^\alpha}(v); \end{aligned}$$

and

$$\begin{aligned} \frac{\partial E}{\partial h^\alpha}(v, h) &= f_\alpha(h) + h^\beta \frac{\partial f_\beta}{\partial h^\alpha}(h) - \frac{\partial L_h}{\partial h^\alpha}(h) - U^{\gamma\beta} g_\gamma(v) \frac{\partial f_\beta}{\partial h^\alpha}(h) \\ \left\{ f_\alpha = \frac{\partial L_h}{\partial h^\alpha} \right\} &= -[U^{\gamma\beta} g_\gamma(v) + h^\beta] \frac{\partial f_\beta}{\partial h^\alpha}(h). \end{aligned}$$

If f and g are piecewise linear functions, then $\partial^2 f$ and $\partial^2 g$ vanish almost everywhere⁷. Thus, comparing with 8, we find $K_v = \partial^2 L_v(v)^{-1}$, $K_h = \partial^2 L_h(h)^{-1}$, and $\nabla K = 0$. That is,

$$\begin{aligned} \frac{dv^\alpha}{dt} &= K_v^{\alpha\beta}(v) [U_{\beta\gamma} f^\gamma(h) - v_\beta + I_\beta] + \sqrt{2T} dW_v^\alpha; \\ \frac{dh^\alpha}{dt} &= K_h^{\alpha\beta}(h) [U_{\gamma\beta} g^\gamma(v) - h_\beta] + \sqrt{2T} dW_h^\alpha, \end{aligned}$$

Thus proof ends. \square

Remark 15. [Hebbian Rule]

In addition, we find, along the gradient descent trajectory of U , the difference is

$$\Delta U^{\alpha\beta} \propto \left\langle -\frac{\partial E}{\partial U^{\alpha\beta}}(v, h) \right\rangle_{q_E(\cdot|v)} = \langle g^\alpha(v) f^\alpha(h) \rangle_{q_E(\cdot|v)}.$$

Since f and g are activation functions, we recover the Hebbian rule, that is, neurons that fire together wire together.

Remark 16. [Simplified Brain]

This model can be viewed as a simplified brain when f and g are linear. Indeed, in the equation (1) of Dehaene et al. (2003)⁸, when the V are limited to a small region, and the τ s are large, then the coefficients, i.e. the m s and h s, can be regarded as constants. The equation (1), thus, reduces to the continuous Hopfield network (without latent variables).

Appendix A Useful Lemmas

Lemma 17. [Kramers–Moyal Expansion]

Given random variable X and time parameter t , consider random variable ϵ whose distribution is (x, t) -dependent. After Δt , particles in position x jump to $x + \epsilon$. Then, we have

$$p(x, t + \Delta t) - p(x, t) = \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!} \nabla_{a_1} \cdots \nabla_{a_n} [p(x, t) M^{a_1 \cdots a_n}(x, t)],$$

where $M^{a_1 \cdots a_n}(x, t)$ represents the n -order moments of ϵ

$$M^{a_1 \cdots a_n}(x, t) := \langle \epsilon^{a_1} \cdots \epsilon^{a_n} \rangle_\epsilon.$$

Proof. The trick is introducing a smooth test function, $h(x)$. Denote

$$I_{\Delta t}[h] := \int d\mu(x) p(x, t + \Delta t) h(x). \quad \square$$

The transition probability from x at t to y at $t + \Delta t$ is $\int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) \delta(x + \epsilon - y)$. This implies

$$p(y, t + \Delta t) = \int d\mu(x) p(x, t) \left[\int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) \delta(x + \epsilon - y) \right].$$

6. Here the $\partial^2 L$ is the Hessian matrix, and $[\partial^2 L]^{-1}$ the inverse matrix.

7. **TODO: Check this.**

8. A neuronal network model linking subjective reports and objective physiological data during conscious perception, Stanislas Dehaene, Claire Sergent, and Jean-Pierre Changeux, 2003.

With this,

$$\begin{aligned}
I_{\Delta t}[h] &:= \int d\mu(x) p(x, t + \Delta t) h(x) \\
\{x \rightarrow y\} &= \int d\mu(y) p(y, t + \Delta t) h(y) \\
[p(y, t + \Delta t) = \dots] &= \int d\mu(x) p(x, t) \int d\mu(y) \int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) \delta(x + \epsilon - y) h(y) \\
\{\text{Integrate over } y\} &= \int d\mu(x) p(x, t) \int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) h(x + \epsilon).
\end{aligned}$$

Taylor expansion $h(x + \epsilon)$ on ϵ gives

$$I_{\Delta t}[h] = \int d\mu(x) p(x, t) h(x) + \sum_{n=1}^{+\infty} \frac{1}{n!} \int d\mu(x) p(x, t) [\nabla_{a_1} \dots \nabla_{a_n} h(x)] \int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) \epsilon^{a_1} \dots \epsilon^{a_n}.$$

Integrating by part on x for the second term, we find

$$I_{\Delta t}[h] = \int d\mu(x) p(x, t) h(x) + \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!} \int d\mu(x) h(x) \nabla_{a_1} \dots \nabla_{a_n} \left[p(x, t) \int d\mu(\epsilon) p_\epsilon(\epsilon; x, t) \epsilon^{a_1} \dots \epsilon^{a_n} \right].$$

Denote n -order moments of ϵ as $M^{a_1 \dots a_n}(x, t) := \langle \epsilon^{a_1} \dots \epsilon^{a_n} \rangle_\epsilon$ and recall the definition of $I_{\Delta t}[h]$, then we arrive at

$$\int d\mu(x) [p(x, t + \Delta t) - p(x, t)] h(x) = \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!} \int d\mu(x) h(x) \nabla_{a_1} \dots \nabla_{a_n} [p(x, t) M^{a_1 \dots a_n}(x, t)].$$

Since $h(x)$ is arbitrary, we conclude that

$$p(x, t + \Delta t) - p(x, t) = \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!} \nabla_{a_1} \dots \nabla_{a_n} [p(x, t) M^{a_1 \dots a_n}(x, t)].$$

Appendix B Stochastic Dynamics

B.1 Random Walk

Given $\forall x \in \mathcal{M}$ and any time t , consider a series of i.i.d. random variables (random walks),

$$\{\varepsilon_i^a; i = 1 \dots n(t)\},$$

where, for $\forall i$, $\varepsilon_i^a \sim P$ for some distribution P , with the mean 0 and covariance $\Sigma(x, t)$, and the walk steps

$$n(t) = \int_0^t d\tau \frac{dn}{d\tau}(x(\tau), \tau).$$

For any time interval Δt , this series of random walks leads to a difference

$$\Delta x^a := \sum_{i=n(t)}^{n(t+\Delta t)} \varepsilon_i^a.$$

Let

$$\tilde{W}^a(x, t) := \frac{1}{\sqrt{n(t+\Delta t) - n(t)}} \sum_{i=n(t)}^{n(t+\Delta t)} \varepsilon_i^a,$$

we have $\Delta x^a = \sqrt{n(t+\Delta t) - n(t)} \tilde{W}^a(x, t)$. Since $n(t+\Delta t) - n(t) = \frac{dn}{dt}(x, t) \Delta t + o(\Delta t)$, we have

$$\Delta x^a = \sqrt{n(t+\Delta t) - n(t)} \tilde{W}^a(x, t) = \sqrt{\frac{dn}{dt}(x, t) \Delta t} \tilde{W}^a(x, t) + o(\sqrt{\Delta t}).$$

If

$$\frac{dn}{dt}(x, t) \Sigma^{ab}(x, t) = \mathcal{O}(1)$$

as $dn/dt \rightarrow +\infty$, that is, more steps per unit time, then, by central limit theorem (for multi-dimension),

$$\Delta x^a = \Delta W^a + o\left(\frac{dn}{dt}(x, t)\right),$$

where

$$\Delta W^a \sim \mathcal{N}(0, \Delta t \Sigma^{ab}(x, t)).$$

B.2 Stochastic Dynamics

TODO