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*

$$\langle f \rangle_p = \langle f \rangle_X = \mathbb{E}_{x\sim p}[f(x)] := \int_{\mathcal{M}} \mathrm{d}\mu(x)\, p(x)\, f(x);$$

- *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;*
- *for conditional maps f, let $f(x|y)$ denotes the map of x with y given and fixed, and $f(x;y)$ denotes the map of x with y given but mutable;*
- *r.v. is short for random variable, and i.i.d. for independent identically distributed.*
- *ODE for ordinary differential equation(s), SDE for stochastic differential equation(s).*
- *Laplacian $\Delta := \nabla_a \nabla^a$.*

1 Lyapunov Function

Definition 2. *[Lyapunov Function]*

Given an autonomous[1] *ODE,*

$$\frac{\mathrm{d}x^a}{\mathrm{d}t} = f^a(x),$$

a Lyapunov function, $V(x)$, of it is a scalar function such that $\nabla_a V(x)\, f^a(x) \leqslant 0$ and the equality holds if and only if $f^a(x) = 0$.

Along the phase trajectory, a Lyapunov function monomotically decreases. So, it reflects the stability of the ODE.

Question 1. Given an autonomous ODE, whether a Lyapunov function of it exists or not?

Question 2. And how to construct, or approximate to, it if there is any?

Here we propose a simulation based method that furnishes a criterion on whether a Lyapunov function exists or not, and then reveals an analytic approximation to the Lyapunov function if it exists.

We first extend the autonomous ODE to a SDE[2], as

$$\mathrm{d}X^a = f^a(X)\, \mathrm{d}t + \sqrt{2T}\, \mathrm{d}W^a,$$

where $\mathrm{d}W^a \sim \mathcal{N}(0, \delta^{ab}\mathrm{d}t)$ and parameter $T > 0$. Then, we sample an essemble of "particles" independently evolving along this SDE. As a set of Markov chains, this simulation will arrive at a stationary distribution. This is true if the Markov chain is irreducible and recurrent. These conditions are hard to check. But, in practice, there is criterion on the convergence of a chain at a finite time.[3] If it has converged, we get an empirical distribution, denoted as p_D, that approximates to the true stationary distribution.

Next, we are to find an analytic approximation to the empirical distribution p_D. This can be taken by any universal approximator, such as neural network. Say, an universal approximator $E(\cdot;\theta)$ parameterized by θ, and define q_E as

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

where $Z_E(\theta) := \int_{\mathcal{M}} \mathrm{d}\mu(x) \exp(-E(x;\theta)/T)$. Then, we construct the loss as

$$L(\theta) := T D_{\mathrm{KL}}(p_D \| q_E(\cdot;\theta)) = T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x) \ln p_D(x) - T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x) \ln q_E(x;\theta).$$

1. That is, ordinary differential equations that do not explicitly depend on time. The word autonomous means independent of time.
2. SDE is defined in ?.
3. E.g., Gelman-Rubin-Brooks plot.

The first term is independent of θ, thus omitable. Thus, the loss becomes

$$\begin{aligned} L(\theta) &= -T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x) \ln q_E(x;\theta) \\ &= \int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x)\, E(x;\theta) + T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x) \ln Z_E(\theta) \\ &= \langle E(\cdot;\theta)\rangle_{p_D} \\ \left[\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p_D(x) = 1\right] &+ T \ln Z_E(\theta). \end{aligned}$$

We find the best fit $\theta_\star := \operatorname{argmin}_\theta L(\theta)$ by using gradient descent. Notice the relation

Lemma 3.

$$T\frac{\partial}{\partial\theta^\alpha} \ln Z_{E(\cdot;\theta)} = -\left\langle \frac{\partial E}{\partial\theta^\alpha}(\cdot;\theta)\right\rangle_{q_E(\cdot;\theta)}.$$

Proof. Directly,

$$\begin{aligned} T\frac{\partial}{\partial\theta^\alpha} \ln Z_{E(\cdot;\theta)} &= T\frac{1}{Z_{E(\cdot;\theta)}}\frac{\partial}{\partial\theta^\alpha} Z_{E(\cdot;\theta)} \\ \{Z_E := \cdots\} &= T\frac{1}{Z_{E(\cdot;\theta)}}\frac{\partial}{\partial\theta^\alpha}\int_{\mathcal{M}} \mathrm{d}\mu(x)\, \mathrm{e}^{-E(x;\theta)/T} \\ &= -\int_{\mathcal{M}} \mathrm{d}\mu(x)\, \frac{\mathrm{e}^{-E(x;\theta)/T}}{Z_{E(\cdot;\theta)}} \frac{\partial E}{\partial\theta^\alpha}(x;\theta) \\ \{q_E := \cdots\} &= -\int_{\mathcal{M}} \mathrm{d}\mu(x)\, q_E(x;\theta) \frac{\partial E}{\partial\theta^\alpha}(x;\theta). \\ &= -\left\langle \frac{\partial E}{\partial\theta^\alpha}(\cdot;\theta)\right\rangle_{q_E(\cdot;\theta)}. \end{aligned}$$

Thus, proof ends. □

This implies

$$\frac{\partial L}{\partial\theta^\alpha}(\theta) = \left\langle \frac{\partial E}{\partial\theta^\alpha}(\cdot;\theta)\right\rangle_{p_D} - \left\langle \frac{\partial E}{\partial\theta^\alpha}(\cdot;\theta)\right\rangle_{q_E(\cdot;\theta)}.$$

Both of the two terms can be computed by Monte Carlo integral. Since p_D has been an empirical distribution, the computation of the first Monte Carlo integral is straight forward. The second can be computed in the same way of generating the empirical distribution p_D, by noticing

Lemma 4. *Markov chains by SDE*

$$\mathrm{d}X^a = -\nabla^a E(x)\, \mathrm{d}t + \sqrt{2T}\, \mathrm{d}V^a,$$

where $\mathrm{d}V^a \sim \mathcal{N}(0, \delta^{ab}\, \mathrm{d}t)$ *and* $T > 0$, *will converge to* q_E.

Proof. By lemma 10, the distribution $p(x,t)$ of the Markov chains generated by the SDE obeys

$$\frac{\partial p}{\partial t}(x,t) = \nabla_a[p(x,t)\, \nabla^a E(x)] + T\, \Delta p(x,t).$$

It's straight forward to check that q_E is a stationary solution to this equation. And for any initial value of $p(x,t)$, it always relax to q_E. Indeed,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} T D_{\mathrm{KL}}(p \| q_E) &= \frac{\mathrm{d}}{\mathrm{d}t} T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p(x,t)\, [\ln p(x,t) - \ln q_E(x)] \\ &= T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, \frac{\partial p}{\partial t}(x,t)\, [\ln p(x,t) - \ln q_E(x) + 1] \\ \left\{\frac{\partial p}{\partial t}(x,t) = \cdots\right\} &= T\int_{\mathcal{M}} \mathrm{d}\mu(x)\, \nabla_a[p(x,t)\, \nabla^a E(x) + T\, \nabla^a p(x,t)]\, [\ln p(x,t) - \ln q_E(x) + 1] \\ \{\text{Integral by part}\} &= -T\int_{\mathcal{M}} \mathrm{d}\mu(x) [p(x,t)\, \nabla^a E(x) + T\, \nabla^a p(x,t)]\, \nabla_a[\ln p(x,t) - \ln q_E(x) + 1] \\ &= -\int_{\mathcal{M}} \mathrm{d}\mu(x)\, p(x,t)\, \nabla_a[E(x) + T \ln p(x,t)]\, \nabla^a[E(x) + T \ln p(x,t)] \\ &\leqslant 0, \end{aligned}$$

and the equility holds if and only if $\nabla_a[E(x) + T \ln p(x,t)] = 0$ for $\forall x$, that is, $p(x,t) \equiv q_E(x)$. □

Even though the θ is keep changing during the gradient descent process, as long as it's controlled so as to be slowly varying, we can use the same strategy as the persistent contrastive divergence trick to simplify the computation. So, during the gradient descent steps, we employ two distinct sets of Markov chains that are consistently evolving. The first is generated by the SDE to p_D, and the second by the SDE to q_E. Along the gradient descent steps of θ, on the chains to p_D E is suppressed, while on the chains to q_E E is elevated. Gradient descent stops when the two parts balance, where q_E fits p_D best.

Finally, we claim that the E we find at the best fit $\theta_\star$ is a Lyapunov function of the original autonomous ODE. By lemma 10, the distribution $p(x,t)$ of the Markov chains generated by the SDE to p_D obeys

$$\frac{\partial p}{\partial t}(x,t) = -\nabla_a[p(x,t)\, f^a(x)] + T\,\Delta p(x,t).$$

In the end, $p \to p_D \approx q_E$ where $\partial p / \partial t \to 0$. Here, it becomes

$$0 = -\nabla_a E(x)\, f^a(x) - \delta^{ab}\,\nabla_a E\,(x) \nabla_b E(x) + T\,[\nabla_a f^a(x) + \Delta E(x)].$$

As $T \to 0$, the SDE reduces to the original autonomous ODE, and we arrive at

$$\nabla_a E(x)\, f^a(x) = -\delta^{ab}\,\nabla_a E\,(x) \nabla_b E(x) \leqslant 0,$$

where equality holds if and only if $\nabla_a E(x) = 0$. Thus, $E(x)$ is a Lyapunov function of $\mathrm{d}x^a / \mathrm{d}t = f^a(x)$.

2 Minimize Free Energy Principle

Definition 5. *[Free Energy]*

Given $E\colon \mathcal{M} \to \mathbb{R}$ and $T > 0$, free energy is a functional of distribution, defined as

$$F_E[p] := \langle E \rangle_p - T H[p].$$

Free energy is a balanced result. Indeed, on one hand, we want to minimize the expectation of energy; and on the other hand, we want to minimize the prior knowledge on p. The parameter T weighs this balance.

Lemma 6. *Let*

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

where $Z_E := \int_{\mathcal{M}} \mathrm{d}\mu(x) \exp(-E(x)/T)$, then we have

$$F[q_E] = -T \ln Z_E.$$

Proof. Since

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

we have

$$\begin{aligned} T H[q_E] &= -T \int_{\mathcal{M}} \mathrm{d}\mu(x)\; q_E(x) \ln q_E(x) \\ &= -T \int_{\mathcal{M}} \mathrm{d}\mu(x)\; q_E(x)\, [-E(x)/T - \ln Z_E] \\ &= \langle E \rangle_{q_E} + T \ln Z_E. \end{aligned}$$

Then,

$$\begin{aligned} F[q_E] &:= \langle E \rangle_{q_E} - T H[q_E] \\ &= \langle E \rangle_{q_E} - \langle E \rangle_{q_E} - T \ln Z_E \\ &= -T \ln Z_E. \end{aligned}$$

Thus proof ends. □

Appendix A Useful Lemmas

A.1 Kramers–Moyal Expansion

Kramers–Moyal Expansion relates the microscopic landscape, i.e. the dynamics of Brownian particles, and the macroscopic landscape, i.e. the evolution of distribution.

Lemma 7. *[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

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

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

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

With this,

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

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

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

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

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

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

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

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

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

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),

$$\{\varepsilon_i^a\colon i=1\ldots n(t)\},$$

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

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

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

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

Then, we have

Theorem 8. *[Brownian Motion]*
As $\mathrm{d}n/\mathrm{d}t\to+\infty$,

$$\Delta x^a=\Delta W^a+o\left(\frac{\mathrm{d}n}{\mathrm{d}t}(x,t)\right),$$

where

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

Proof. 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 $n(t+\Delta t)-n(t)=\frac{\mathrm{d}n}{\mathrm{d}t}(x,t)\,\Delta t+o(\Delta t)$, we have

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

If

$$\frac{\mathrm{d}n}{\mathrm{d}t}(x,t)\,\Sigma^{ab}(x,t)=\mathcal{O}(1)$$

as $\mathrm{d}n/\mathrm{d}t\to+\infty$, that is, more steps per unit time, then, by central limit theorem (for multi-dimension),

$$\Delta x^a=\Delta W^a+o\left(\frac{\mathrm{d}n}{\mathrm{d}t}(x,t)\right),$$

where

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

□

B.2 Stochastic Dynamics

A stochastic dynamics, or stochastic differential equations (SDE), is defined by two parts. The first is deterministic, and the second is a random walk. Precisely,

Definition 9. *Given* $f^a(x,t)$, $g_b^a(x,t)$, *and* $\Sigma^{ab}(x,t)$ *on* $\mathcal{M}\times\mathbb{R}$, *stochastic differential equations is defined as*

$$\mathrm{d}x^a=f^a(x,t)\,\mathrm{d}t+g_b^a(x,t)\,\mathrm{d}W^b(x,t),$$

where $\mathrm{d}W^a(x,t)$ *is a random walk with covariance* $\Sigma^{ab}(x,t)\,\mathrm{d}t$.

Lemma 10. *[Macroscopic Landscape]*
Consider an ensemble of particles, randomly sampled at an initial time, evolving along a SDE ?. By saying "ensemble", we mean that the number of particles has the order of Avogadro's constant, s.t. the distribution of the particles can be viewed as smooth. Let $p(x,t)$ *denotes the distribution. Then we have*

$$\frac{\partial p}{\partial t}(x,t)=-\nabla_a[p(x,t)\,f^a(x,t)]+\frac{1}{2}\nabla_a\nabla_b[p(x,t)\,K^{ab}(x,t)],$$

where $K^{ab}:=g_c^a(x,t)\,g_d^b(x,t)\,\Sigma^{cd}(x,t)$.

Proof. From the difference of the SDE,

$$\Delta x^a=f^a(x,t)\,\Delta t+g_b^a(x,t)\,\Delta W^a(x,t),$$

by Kramers–Moyal expansion 7, 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 $\mathrm{d}W^a(x,t)$ is a random walk, $\langle\Delta W^a(x,t)\rangle_{\Delta W(x,t)}=0$. Then the term is

$$-\nabla_a[p(x,t)\langle\Delta x^a\rangle_{\Delta x}]=-\nabla_a[p(x,t)\,f^a(x,t)]\Delta t.$$

And for $n=2$, by noticing that, as a random walk, $\langle\Delta W^a(x,t)\,\Delta W^b(x,t)\rangle_{\Delta W(x,t)}=\mathcal{O}(\Delta t)$, we have,

$$\frac{1}{2}\nabla_a\nabla_b[p(x,t)\,\langle\Delta x^a\,\Delta x^b\rangle_{\Delta x}]=\frac{1}{2}\nabla_a\nabla_b[p(x,t)\,g_c^a(x,t)\,g_d^b(x,t)\,\Sigma^{cd}(x,t)]\,\Delta t+o(\Delta t).$$

For $n\geqslant 3$, all are $o(\Delta t)$. So, we have

$$p(x,t+\Delta t)-p(x,t)=-\nabla_a[p(x,t)\,f^a(x,t)]+\frac{1}{2}\nabla_a\nabla_b[p(x,t)\,g_c^a(x,t)\,g_d^b(x,t)\,\Sigma^{cd}(x,t)]\,\Delta t+o(\Delta t).$$

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

$$\frac{\partial p}{\partial t}(x,t)=-\nabla_a[p(x,t)\,\mu^a(x,t)]+\frac{1}{2}\nabla_a\nabla_b[p(x,t)\,g_c^a(x,t)\,g_d^b(x,t)\,\Sigma^{cd}(x,t)].$$

Thus proof ends. □