Fluctuation Suppression and Enhancement in Interacting Particle Systems

Jiaheng Chen

Shanghai Jiao Tong University

April 7, 2022

Overview

1. Part I : Kernel Stein Discrepancy Descent and its Advantages in Sampling

2. Part II : Fluctuation Suppression and Enhancement in General Interacting Particle Systems

Problem Setting

Problem

Sample from a target distribution π over $\mathbb{R}^d$, whose density w.r.t. Lebesgue is known up to a constant Z:

$$\pi(x) = \frac{\tilde{\pi}(x)}{Z}$$

where Z is the (untractable) normalization constant.

Motivation:

- Let $\mathcal{D} = (w_i, y_i)_{i=1,\cdots,N}$ observed data.
- Assume an underlying model parametrized by θ
  (e.g.$p(y|w,\theta)$ gaussian)
  $\Rightarrow$ Likelihood:$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|w_i,\theta)$.
- Assume also $\theta \sim p$(prior distribution).
  Bayes's rule: $\pi(\theta) := p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{Z}$, $Z = \int_{\mathbb{R}^d} p(\mathcal{D}|\theta)p(\theta)d\theta$.

Sampling as optimization over distributions

- Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$.
- The sampling task can be recast as an optimization problem:

$$\pi = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{argmin} \; D(\mu|\pi) := \mathcal{F}(\mu),$$

  where D is a **dissimilarity functional**.
- Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, one can then consider the **Wasserstein gradient flow** of $\mathcal{F}$ over $\mathcal{P}_2(\mathbb{R}^d)$ to transport μ_0 to π.

Choice of the loss function

Many possibilities for the choice of D among Wasserstein distances, f-divergences, Integral Probability Metrics...

- D is the Kullback-Leibler divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log(\frac{\mu}{\pi}) d\mu & \textit{if } \mu \ll \pi, \\ +\infty & \textit{otherwise}. \end{cases}$$

- D is the MMD (Maximum Mean Discrepancy):

$$\mathrm{MMD}^2(\mu, \pi) = \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu(y) + \iint_{\mathbb{R}^d} k(x,y) d\pi(x) d\pi(y) - 2 \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\pi(y),$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a p.s.d. kernel.

Kernel Stein Discrepancy (Liu et al.2016)[5]

For $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the KSD of μ relative to π is

$$\mathrm{KSD}(\mu|\pi) = \sqrt{\iint k_\pi(x, y) d\mu(x) d\mu(y)},$$

where $k_\pi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the **Stein kernel**, defined through

- **the score function** $s_\pi(x) = \nabla \mathrm{log} \pi(x)$,
- a p.s.d. kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, k \in C^2(\mathbb{R}^d)$. (e.g.$k(x, y) = exp(-\|x - y\|^2/h)$)

For $x, y \in \mathbb{R}^d$,

$$k_\pi(x, y) = s(x)^T s(y) k(x, y) + s(x)^T \nabla_2 k(x, y) + \nabla_1 k(x, y)^T s(y) + \nabla \cdot_1 \nabla_2 k(x, y).$$

Equivalently,

$$\mathrm{KSD}^2(\mu|\pi) = \mathbb{E}_{x,y\sim\mu}\Big[(s_\pi(x) - s_\mu(x))^T k(x, y)(s_\pi(y) - s_\mu(y))\Big].$$

Stein identity and link with MMD

Under mild assumptions on k and π, the Stein kernel k_π is **p.s.d.** and satisfies a **Stein identity**

$$\int_{\mathbb{R}^d} k_\pi(x, \cdot) d\pi(x) = 0.$$

Consequently, **KSD is a MMD** with kernel k_π, since:

$$\begin{aligned}
\mathrm{MMD}^2(\mu|\pi) &= \int k_\pi(x, y) d\mu(x) d\mu(y) + \int k_\pi(x, y) d\pi(x) d\pi(y) \\
&\quad - 2 \int k_\pi(x, y) d\mu(x) d\pi(y) \\
&= \int k_\pi(x, y) d\mu(x) d\mu(y) \\
&= \mathrm{KSD}^2(\mu|\pi).
\end{aligned}$$

KSD benefits

KSD can be computed when

- one has access to the score of π,
- μ is a discrete measure, e.g. $\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}$, then

$$\mathrm{KSD}^2(\mu|\pi) = \frac{1}{N^2}\sum_{i,j=1}^{N} k_\pi(x^i, x^j).$$

KSD is known to metrize weak convergence [2] when:

- π is strongly log-concave at infinity,
- k has a slow decay rate.

KSD in the literature

The KSD has been used for

- nonparametric statistical tests for goodness-of-fit [Xu and Matsuda, 2020, Kanagawa et al.,2020]
- sampling tasks
  - (greedy algorithms) to select a suitable set of static points to approximate π, adding a new one at each iteration, [Chen et al.,2018, Chen et al.,2019]
  - to compress [Riabiz et al.,2020] or reweight [Hodgkinson et al., 2020] Markov Chain Monte Carlo (MCMC) outputs,
  - to learn a static transport map from μ_0 to π [Fisher et al., 2020],
  - to learn Energy-Based models $\pi \propto \exp(-V)$ from samples of π (use reverse KSD) [Domingo Enrich et al.,2021].

Time/Space discretization of the KSD gradient flow

Let $\mathcal{F}(\mu) = \mathrm{KSD}^2(\mu|\pi)$.

- Its Wasserstein gradient flow on $\mathcal{P}_2(\mathbb{R}^d)$ finds a continuous path of distributions that decreases $\mathcal{F}$.
- Different algorithms to approximate π depend on the time and space discretization of this flow.

Forward discretization: Wasserstein gradient descent
Discrete measures: For discrete measures $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$, we have an explicit loss function

$$L([x^i]_{i=1}^N) := \mathcal{F}(\hat{\mu}) = \frac{1}{N^2}\sum_{i,j=1}^{N} k_\pi(x^i, x^j).$$

Then, Wasserstein gradient descent of $\mathcal{F}$ for discrete measures

$$\Updownarrow$$

(Euclidean) gradient descent of L on the particles.

KSD Descent – algorithms (Korba et al.2021) [4]

One direct way to implement KSD Descent (Gradient descent):

Algorithm 1 KSD Descent GD

Input: initial particles $(x_0^i)_{i=1}^N \sim \mu_0$, number of iterations M, step-size γ

for $n = 1$ **to** M **do**

$$[x_{n+1}^i]_{i=1}^N = [x_n^i]_{i=1}^N - \frac{\gamma}{N^2} \sum_{j=1}^{N} [\nabla_2 k_\pi(x_n^j, x_n^i)]_{i=1}^N, \tag{12}$$

end for

Return: $[x_M^i]_{i=1}^N$.

KSD Descent as interacting particle system

- KSD Descent is a sampling algorithm based on the following interacting particle systems (after time scaling)

$$\begin{cases} \dot{X}_i = -\frac{1}{N} \sum_{j=1}^{N} \nabla k_\pi(X_i, X_j) \\ \{X_i(0)\}_{i=1}^{N} \sim \mu_0 \end{cases}$$

- The empirical measure $\mu_N := \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i(t))$, then

$$\dot{X}_i = -\int \nabla k_\pi(X_i, x)\mu_N(dx) = -\nabla\Big(\int k_\pi(X_i, x)\mu_N(dx)\Big).$$

- Generally, consider the following ODE system of $\{X_i\}_{i=1}^{N}$

$$\dot{X}_i = -\nabla V(X_i, \mu_N), \quad i = 1, \cdots, N.$$

Our interests and motivation

$$\dot{X}_i = -\nabla V(X_i, \mu_N) \quad \rightsquigarrow \quad \partial_t \mu_N = \nabla \cdot (\nabla V(x, \mu_N)\mu_N).$$

- As $N \to \infty$, μ_N can be shown to converge in some sense to the Fokker-Planck equation [6]

$$\partial_t \mu = \nabla \cdot (\nabla V(x, \mu)\mu).$$

- Now suppose that μ_N converges to μ, the **fluctuation** in the $N \to \infty$ limit

$$\eta := \lim_{N\to\infty} \sqrt{N}(\mu_N - \mu).$$

- Question: How will the fluctuation evolve during the dynamics?
  - If the particle are i.i.d. sampled, the fluctuation follows the Central Limit Theorem (CLT) and has variance $1/N$.
  - No longer simple since the dynamics introduce interactions among the particles.

Our interests and motivation

- The interacting particle system:

$$\dot{X}_i = -\nabla V(X_i, \mu_N) \quad \rightsquigarrow \quad \partial_t \mu_N = \nabla \cdot (\nabla V(x, \mu_N)\mu_N). \tag{1}$$

- The mean field equation

$$\dot{X}_i = -\nabla V(X_i, \mu) \quad \rightsquigarrow \quad \partial_t \mu = \nabla \cdot (\nabla V(x, \mu)\mu). \tag{2}$$

- If there are N particles drawn $\bar{X}_i(0)$ i.i.d. from ρ_0 and they evolve according to the ODE (2), then they will be independent from each other for any $t > 0$.
- Then these particles can be viewed as the Monte Carlo samplings from ρ for every t. The fluctuation in this case

$$\bar{\eta} := \lim_{N \to \infty} \sqrt{N}(\bar{\mu}_N - \mu).$$

- We will compare $\|\eta_t\|$ with $\|\bar{\eta}_t\|$, here $\|\cdot\|$ is some norm.

Flow mapping methods in Chen et. al.2020[1]

- The mean field Wasserstein gradient flow

$$\partial_t \mu_t = \nabla \cdot (\nabla V(x, \mu_t)\mu_t), \quad \mu_{t=0} = \mu_0. \tag{3}$$

- Interpreted as the pushforward of the characteristic flow map

$$\int \chi(x)\mu_t(dx) = \int \chi(\Theta_t(x))\mu_0(dx),$$

where χ is a continuous test function and Θ_t solves

$$\dot{\Theta}_t(x) = -\nabla V(\Theta_t(x), \mu_t), \quad \Theta_0(x) = x.$$

- Similarly, for Wasserstein gradient flow of the empirical measure

$$\dot{\Theta}_t^{(N)}(x) = -\nabla V(\Theta_t^{(N)}(x), \mu_t^{(N)}), \quad \Theta_0^{(N)}(x) = x.$$

Flow mapping methods in [1]

- $\eta_t^{(N)} := \sqrt{N}(\mu_t^{(N)} - \mu_t)$
- Take a test function $\chi(x)$,

$$
\begin{aligned}
\int \chi(x)\eta_t^{(N)}(dx) &= \sqrt{N}\int \chi(x)\Big(\mu_t^{(N)}(dx) - \mu_t(dx)\Big) \\
&= \sqrt{N}\int \chi(\Theta_t^{(N)}(x))\mu_0^{(N)}(dx) - \chi(\Theta_t(x))\mu_0(dx) \\
&= \sqrt{N}\int \chi(\Theta_t^{(N)}(x))\mu_0^{(N)}(dx) - \chi(\Theta_t(x))\mu_0^{(N)}(dx) \\
&\qquad + \chi(\Theta_t(x))\mu_0^{(N)}(dx) - \chi(\Theta_t(x))\mu_0(dx) \\
&= \int \chi(\Theta_t(x))\eta_0^{(N)}(dx) + \sqrt{N}\Big[\chi(\Theta_t^{(N)}(x)) - \chi(\Theta_t(x))\Big]\mu_0^{(N)}.
\end{aligned}
$$

- **The first term**: $\Theta_t^{(N)}$ remains equal to Θ_t.
- **The second term** captures the deviation to the flow Θ_t induced by the perturbation of μ_0, i.e. **how much** $\Theta_t^{(N)}$ **differs from** Θ_t.

Flow mapping methods in [1]

Proposition 3.1[1]

Under mild conditions, $\forall t > 0$, as $N \to \infty$ we have $\eta_t^{(N)} \rightharpoonup \eta_t$ weakly in law with respect to $\mathbb{P}_0$, where η_t is such that given a test function χ,

$$\int \chi(x)\eta_t(dx) = \int \chi(\Theta_t(x))\eta_0(dx) + \int \nabla\chi(\Theta_t(x)) \cdot T_t(x)\mu_0(dx).$$

Here η_0 is the Gaussian measure with mean zero and covariance

$$\mathbb{E}_0[\eta_0(dx)\eta_0(dx')] = \mu_0(dx)\delta_x(dx') - \mu_0(dx)\mu_0(dx'),$$

and $T_t = \lim_{N\to\infty} \sqrt{N}(\Theta_t^{(N)} - \Theta_t)$ is the flow solution to

$$\dot{T}_t(x) = -\nabla\nabla V(\Theta_t(x), \mu_t)T_t(x) - \int \nabla K(\Phi_t(x), x')\eta_t(dx')$$

Note: This proposition holds for $V(x, \mu) = F(x) + \int K(x, x')\mu(dx')$.

The fluctuation in KSD Descent

- KSD Descent:

$$\dot{X}_i = -\int \nabla k_\pi(X_i, x')\mu_N(dx') = -\nabla\Big(\int k_\pi(X_i, x')\mu_N(dx')\Big)$$

where

$$\begin{aligned} k_\pi(x, x') = s_\pi(x) \cdot s_\pi(x')k(x, x') + s_\pi(x) \cdot \nabla' k(x, x') \\ + \nabla k(x, x') \cdot s_\pi(x') + \mathrm{tr}(\nabla\nabla' k(x, x')) \end{aligned}$$

and $s_\pi(x) = \nabla \log \pi(x)$.

- KSD Descent can be seen as a specific example when

$$V(x, \mu) = \int k_\pi(x, x')\mu(dx').$$

The fluctuation in KSD Descent

- Recall that

$$\int \chi(x)\eta_t(dx) = \int \chi(\Theta_t(x))\eta_0(dx) + \int \nabla\chi(\Theta_t(x)) \cdot T_t(x)\mu_0(dx)$$

  where T_t is the flow solution to

$$\dot{T}_t(x) = -\nabla\nabla V(\Theta_t(x), \mu_t)\, T_t(x) - \int \nabla k_\pi(\Theta_t(x), x')\eta_t(dx').$$

- By the Duhamel's principle

$$T_t(x) = -\int_0^t J_{t,s}(x) \int \nabla k_\pi(\Theta_s(x), x')\eta_s(dx')ds,$$

  where $J_{t,s}$ is the solution to

$$\frac{d}{dt} J_{t,s}(x) = -\nabla\nabla V(\Theta_t(x), \mu_t) J_{t,s}(x), \quad J_{s,s}(x) = Id.$$

The fluctuation in KSD Descent

Theorem 3.7[5]

Assume $k(x, x')$ is a positive definite kernel with positive eigenvalues $\{\lambda_j\}$ and eigenfunctions $\{e_j(x)\}$, then $k_\pi(x, x')$ is also a positive definite kernel, and can be rewritten into

$$k_\pi(x, x') = \sum_j \lambda_j [\mathcal{A}_\pi e_j(x)]^T [\mathcal{A}_\pi e_j(x')],$$

where $\mathcal{A}_\pi e_j(x) = s_\pi(x) e_j(x) + \nabla e_j(x)$ is the Stein's operator acted on e_j. In addition,

$$\mathrm{KSD}^2(\mu|\pi) = \mathbb{E}_{x,x' \sim \mu} k_\pi(x, x') = \sum_j \lambda_j \|\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi e_j(x)]\|_2^2.$$

The fluctuation in KSD Descent

- Calculations:

$$T_t(x) = -\int_0^t J_{t,s}(x) \int \nabla k_\pi(\Theta_s(x), x')\eta_s(dx')ds$$
$$= -\sum_i \lambda_i \int_0^t J_{t,s}(x)\nabla \mathcal{A}_\pi e_i(\Theta_s(x)) \int \mathcal{A}_\pi e_i(x')\eta_s(dx')ds.$$

- Introduce $g_t^{(j)} := \int \mathcal{A}_\pi e_j(x')\eta_t(dx')$.
- By the property of η_t:

$$g_t^{(j)} = \int \mathcal{A}_\pi e_j(\Theta_t(x))\eta_0(dx) + \int \nabla \mathcal{A}_\pi e_j(\Theta_t(x)) \cdot T_t(x)\mu_0(dx)$$
$$= \bar{g}_t^{(j)} - \sum_i \lambda_i \int_0^t \Gamma_{t,s}^{i,j}\, g_s^{(i)} ds$$

where

$$\Gamma_{t,s}^{i,j} = \int \nabla \mathcal{A}_\pi e_j(\Theta_t(x)) J_{t,s}(x) \nabla \mathcal{A}_\pi e_i(\Theta_s(x))\mu_0(dx).$$

The fluctuation in KSD Descent

For every j it holds that

$$g_t^{(j)} = \bar{g}_t^{(j)} - \sum_i \lambda_i \int_0^t \Gamma_{t,s}^{i,j} \, g_s^{(i)} ds.$$

Taking the dot product by $\lambda_j g_t^{(j)}$ on both sides and sum over j

$$\sum_j \lambda_j |g_t^{(j)}|^2 = \sum_j \lambda_j g_t^{(j)} \cdot \bar{g}_t^{(j)} - \sum_{i,j} \lambda_i \lambda_j \int_0^t \langle g_t^{(j)}, \Gamma_{t,s}^{i,j} \, g_s^{(i)} \rangle.$$

Let $\phi(t,x) := \sum_j \lambda_j \nabla \mathcal{A}_\pi e_j(\Theta_t(x)) g_t^{(j)}$, then

$$\sum_j \lambda_j |g_t^{(j)}|^2 = \sum_j \lambda_j g_t^{(j)} \cdot \bar{g}_t^{(j)} - \int_0^t \int \langle \phi(t,x), J_{t,s}(x) \phi(s,x) \rangle \mu_0(dx) ds.$$

The fluctuation in KSD Descent

$$\sum_j \lambda_j |g_t^{(j)}|^2 = \sum_j \lambda_j g_t^{(j)} \cdot \bar{g}_t^{(j)} - \int_0^t \int \langle \phi(t,x), J_{t,s}(x)\phi(s,x)\rangle \mu_0(dx)ds.$$

- $J_{t,s}$ satifies

$$\frac{d}{dt} J_{t,s}(x) = -\nabla\nabla V(\Theta_t(x), \mu_t) J_{t,s}(x), \quad J_{s,s}(x) = Id.$$

- If $J_{t,s}$ is a nonnegative Volterra kernel, then for every $T > 0$

$$\int_0^T \sum_j \lambda_j |g_t^{(j)}|^2 dt \le \int_0^T \sum_j \lambda_j g_t^{(j)} \cdot \bar{g}_t^{(j)} dt,$$

which implies that

$$\int_0^T \sum_j \lambda_j |g_t^{(j)}|^2 dt \le \int_0^T \sum_j \lambda_j |\bar{g}_t^{(j)}|^2 dt.$$

Some comments on the fluctuation in KSD Descent

- Under the thermal equilibrium, namely $\mu_0 = \mu_t = \mu_\infty$, $\Theta_t(x) = \Theta_\infty(x) \equiv x$ and $\nabla\nabla V(x, \mu_\infty)$ is p.s.d., then

  $$J_{t,s} = e^{-(t-s)\nabla\nabla V(x,\mu_\infty)}$$

  is a nonnegative Volterra kernel, which means

  $$\int_0^T \int_0^t \langle \phi(t), J(t-s)\phi(s)\rangle dsdt \geq 0.$$

  Then

  $$\int_0^T \sum_j \lambda_j |g_t^{(j)}|^2 dt \leq \int_0^T \sum_j \lambda_j |\bar{g}_t^{(j)}|^2 dt.$$

- Recall:

  $$\mathrm{KSD}^2(\mu|\pi) = \sum_j \lambda_j \Big| \int \mathcal{A}_\pi e_j(x)\mu(x) \Big|^2.$$

  $$g_t^{(j)} = \int \mathcal{A}_\pi e_j(x)\eta_t(dx), \quad \eta_t = \lim_{N\to\infty} \sqrt{N}(\mu_N - \mu).$$

- Here the norm is $\|\eta_t\|_{k_\pi}^2 := \iint k_\pi(x, x')\eta_t(dx)\eta_t(dx')$.

General interacting particle systems

- Generally, the first order SDE systems for N interacting particles in the mean field scaling

$$dX_i = -\nabla V(X_i)dt - \frac{1}{N}\sum_j \nabla W(X_i - X_j)dt + \sqrt{2\beta^{-1}}dB_i, \quad i = 1, \cdots, N.$$

- The corresponding Fokker-Planck equation

$$\partial_t \rho = \nabla \cdot ((\nabla V + \nabla W * \rho)\rho) + \beta^{-1}\Delta\rho.$$

- Note: For the system with noise, the approach in [1] using the flow mapping is not accessible.
- The SPDE that the fluctuation satisfies [7]

$$\partial_t \eta = \nabla \cdot (\nabla U(x,t)\eta) + \beta^{-1}\Delta\eta + \nabla \cdot (\nabla W * \eta\mu_t) - \sqrt{2\beta^{-1}}\nabla \cdot (\sqrt{\mu_t}\xi)$$

where $U(x,t) = V(x) + W * \mu$ and ξ is a space-time noise.

Main Results

The basic equations in the thermal equilibrium

Proposition 1

Both $\hat{\eta}_t$ and $\hat{\bar{\eta}}_t$ are Gaussian stochastic processes. They satisfy the relation

$$\hat{\eta}_t(\omega) = \hat{\bar{\eta}}_t(\omega) \mp \frac{1}{(2\pi)^d} \int_0^t \int_{\hat{\mathbf{X}}} k(\omega, \omega', t-s) \hat{\Phi}(\omega') \hat{\eta}_s(\omega') d\omega' ds,$$

where "$-$" sign corresponds to $W = \Phi$ and "$+$" corresponds to $W = -\Phi$ respectively, and

$$k(\omega, \omega', s) = \beta \int_{\mathbf{X}} \left(e^{-\frac{1}{2}s\mathcal{A}} e^{-i\omega \cdot y} \right) \mathcal{A}(e^{-\frac{1}{2}s\mathcal{A}} e^{i\omega' \cdot y}) \mu_*(dy).$$

Here $\mathcal{A} = -\mathcal{L} = \nabla U(x) \cdot \nabla - \beta^{-1}\Delta = -\beta^{-1} e^{\beta U} \nabla \cdot (e^{-\beta U} \nabla)$. For each s, k is Hermitian with

$$k(\omega, \omega', s) = \overline{k(\omega', \omega, s)}$$

and is positive semi-definite in s.

Main Results

Reduced system using eigen-expansion

- Assume $\mathcal{L}$ has a spectral gap, then $\mathcal{A} = -\mathcal{L}$ is a nonnegative self-adjoint operator in $L^2(\mathbb{R}^d; \mu_*)$ with discrete spectrum. The eigenvalue problem for the generator is

$$-\mathcal{L}\phi_n = \lambda_n \phi_n, \quad n = 0, 1, ...$$

Proposition 2

For all i, j

$$G_{ij} = \iint_{\mathbf{X}\times\mathbf{X}} \Phi(y - y')\phi_i(y)\phi_j(y')\mu_*(dy)\mu_*(dy') \in \mathbb{R}.$$

The operator $G : \ell^2 \to \ell^2$ is positive semi-definite. If moreover $\hat{\Phi}$ has full support in $\hat{X}$, G is positive definite.

Main Results

Reduced system using eigen-expansion

- Introduce $\tilde{X}_i(t) = \int_{\mathbf{X}} \phi_i(y)\eta_t(dy)$, $\quad \tilde{Y}_i(t) = \int_{\mathbf{X}} \phi_i(y)\bar{\eta}_t(dy)$.
- Define $X := G^{1/2}\tilde{X}$, $\quad Y := G^{1/2}\tilde{Y}$.

Proposition 3

(i) Almost surely, $X(t) = G^{1/2}\tilde{X}(t) \in \ell^2$ and $Y(t) = G^{1/2}\tilde{Y}(t) \in \ell^2$.

(ii) It holds that

$$\|\eta_t\|_{\Phi}^2 = \|\hat{\eta}_t\|_{L^2(\nu)}^2 = \langle X, X\rangle_{\ell^2} = \langle \tilde{X}, G\tilde{X}\rangle_{\ell^2}.$$

and similar relations hold for $\bar{\eta}_t$ and $Y(t)$.

(iii) Introducing a family of operators $\Lambda(t) : \ell^2 \to \ell^2$ for $t > 0$, defined by $(\Lambda(t)X)_i = \lambda_i e^{-\lambda_i t} X_i$, then the following equation holds

$$X(t) = Y(t) \mp \beta \int_0^t G^{1/2}\Lambda(t-s)G^{1/2}X(s)ds, \tag{4}$$

where "$-$" sign corresponds to $W = \Phi$ and "$+$" corresponds to $W = -\Phi$ respectively.

Main Results

The space homogenous systems on torus

Theorem 1

1. If $W = \Phi$, $\mathbb{E}\|\eta_t\|_\Phi^2$ is decreasing in time, and for any $t > 0$

$$\mathbb{E}\|\eta_t\|_\Phi^2 < \mathbb{E}\|\bar{\eta}_t\|_\Phi^2.$$

   Moreover, for $j \geq 1$, as $t \to \infty$, one has

$$\lim_{t\to\infty} \mathbb{E}\|\eta_t\|_\Phi^2 = \sum_{j\geq 1} \frac{\mathbb{E}|Y_j|^2}{1 + \beta\mathbb{E}|Y_j|^2},$$

   and consequently $\lim_{\beta\to+\infty} \lim_{t\to\infty} \|\eta_t\|_\Phi^2 = 0$.

2. If $W = -\Phi$, $\mathbb{E}\|\eta_t\|_\Phi^2$ is increasing in time, and for any $t > 0$

$$\mathbb{E}\|\eta_t\|_\Phi^2 > \mathbb{E}\|\bar{\eta}_t\|_\Phi^2,$$

   Moreover, there is a critical value β_c such that when $\beta > \beta_c$, $\lim_{t\to\infty} \mathbb{E}\|\eta_t\|_\Phi^2 = +\infty$.

Main Results

General cases

Lemma 1

With the notations introduced in Proposition 3, it holds almost surely that

$$\|\hat{\eta}_t\|^2_{L^2(\nu)} = \begin{cases} \|\hat{\hat{\eta}}_t\|^2_{L^2(\nu)} + \mathcal{R}_+(t), & \text{if} \quad W = \Phi; \\ \|\hat{\hat{\eta}}_t\|^2_{L^2(\nu)} + \mathcal{R}_-(t), & \text{if} \quad W = -\Phi, \end{cases}$$

where

$$\mathcal{R}_\pm(t) = \mp 2\beta \Big\langle X(t), \int_0^t G^{1/2}\Lambda(t-s)G^{1/2}X(s)ds \Big\rangle_{\ell^2} \\ - \beta^2 \Big\| \int_0^t G^{1/2}\Lambda(t-s)G^{1/2}X(s)ds \Big\|^2_{\ell^2}.$$

Main Results

General cases

Theorem 2

1. ($W = \Phi$, positive definite case) For any $T > 0$, it holds almost surely that

$$\int_0^T \|\eta_t\|_\Phi^2 dt \leq \int_0^T \|\bar{\eta}_t\|_\Phi^2 dt.$$

2. ($W = -\Phi$, negative definite case) Assume the interaction is weak such that

$$\|G\| \leq 2\beta^{-1},$$

   where $\|\cdot\|$ is the operator norm. Then for any $T > 0$ it holds almost surely that

$$\int_0^T \|\eta_t\|_\Phi^2 dt \geq \int_0^T \|\bar{\eta}_t\|_\Phi^2 dt.$$

- Relates to Volterra equation with convolution kernels of positive type[3].
- The condition $\|G\| \leq 2\beta^{-1}$ is equivalent to that $G^{1/2}\Lambda(t-s)G^{1/2}$ is of **anti-coercive type** with coercivity constant $q = 2\beta^{-1}$.

Summary

- **KSD Descent** is a sampling algorithm based on Wasserstein gradient flow and interacting particle system.
- In the equilibrium, KSD Descent introduces **smaller** fluctuation compared with standard Monte Carlo sampling and has better sampling properties.
- Generally, the systems with positive definite interaction potentials tend to exhibit smaller fluctuation compared to the fluctuation in standard Monte Carlo sampling while systems with negative definite potentials tend to exhibit larger fluctuation.

Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden.
A dynamical central limit theorem for shallow neural networks.
Advances in Neural Information Processing Systems, 33:22217–22230, 2020.

Jackson Gorham and Lester Mackey.
Measuring sample quality with kernels.
In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

Gustaf Gripenberg, Stig-Olof Londen, and Olof Staffans.
Volterra integral and functional equations.
Number 34. Cambridge University Press, 1990.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin.
Kernel stein discrepancy descent.
In *International Conference on Machine Learning*, pages 5719–5730. PMLR, 2021.

Qiang Liu, Jason Lee, and Michael Jordan.
A kernelized stein discrepancy for goodness-of-fit tests.
In *International conference on machine learning*, pages 276–284. PMLR, 2016.

Alain-Sol Sznitman.
Topics in propagation of chaos.

Thanks for your listening!