Interacting Particle Systems

Idiosyncratic noise, common noise, and mean-field limits

Hugo Koubbi

Young Researchers Day

Consider particles $(X_1, \dots, X_N) \in (\mathbb{R}^d)^N$ and the mean-field energy

$$H_N(x_1, \dots, x_N) = \sum_{i=1}^{N} V(x_i) + \frac{1}{2N} \sum_{i,j=1}^{N} W(x_i - x_j),$$

where $W, V : \mathbb{R}^d \to \mathbb{R}$.

Equilibrium statistical physics predicts convergence towards the Gibbs measure $\pi_N \in \mathcal{P}((\mathbb{R}^d)^N)$

$$\pi_N(\mathrm{d}x_1, \dots, \mathrm{d}x_N) = Z_N^{-1} \exp\big[-\beta H_N(x)\big]\mathrm{d}x.$$

The dynamics should be chosen so that this measure is invariant.

The overdamped Langevin dynamics associated with H_N is

$$\mathrm{d}X_i(t) = -\nabla_{X_i} H_N(X_i(t))\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B_i(t).$$

where $(B_i)_{i=1}^N$ are independent Brownian motions.

Denote $f_t^N(x_1, \dots, x_N)$ the law of $(X_1(t), \dots, X_N(t))$, it satisfies

$$\partial_t f_t^N(x_1, \dots, x_N) = \sum_{i=1}^N \mathrm{div}_{x_i}(f_t^N \nabla_{x_i} H_N) + \frac{1}{\beta}\sum_{i=1}^N \Delta_{x_i} f_t^N$$

Stationnary measure is given by

$$\pi^N(x_1, \dots, x_N) = \frac{1}{Z_N}\exp\left(-\beta H_N(x_1, \dots, x_n)\right)\mathrm{d}x_1 \dots \mathrm{d}x_n$$

Introduce the empirical measure

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i(t)} \in \mathcal{P}(\mathbb{R}^d).$$

For a smooth test function φ, by Itô formula

$$\begin{aligned}\mathrm{d}\left\langle \mu_t^N, \varphi \right\rangle = \langle \mu_t^N, -(\nabla V + \nabla W * \mu_t^N)\cdot\nabla\varphi + \beta^{-1}\Delta\varphi\rangle \mathrm{d}t \\ + \frac{1}{N}\sqrt{\frac{2}{\beta}}\sum_{i=1}^N \nabla\varphi(X_i(t))\cdot \mathrm{d}B_i(t),\end{aligned}$$

where we denote $\langle \mu, \varphi\rangle = \int \varphi(x)\mu(\mathrm{d}x)$, and $\cdot$ is the euclidean dot product.

$$\frac{2}{N^2\beta}\mathbb{E}[\sum_{i=1}^N \nabla\varphi(X_i(t))\cdot \mathrm{d}B_i(t)^2] = \frac{2}{N^2\beta}\sum_{i=1}^n \|\nabla\varphi(X_i(t))\|^2 \simeq \frac{1}{N}$$

The martingale has quadratic variation of order $1/N$; therefore it disappears for large-N.

The McKean–Vlasov limit

The law solves the nonlinear Fokker–Planck equation

$$\partial_t \mu_t = \operatorname{div}\big(\mu_t(\nabla V + \nabla W * \mu_t)\big) + \beta^{-1}\Delta\mu_t.$$

We can define a limiting nonlinear process with the same law

$$\mathrm{d}\overline{X}_t = -\nabla V(\overline{X}_t)\mathrm{d}t - \nabla W * \mu_t(\overline{X}_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\,\mathrm{d}B_t, \qquad \mu_t = \mathcal{L}(\overline{X}_t).$$

Propagation of chaos

If the initial particles are asymptotically independent, then for fixed time T any finite group of particles becomes asymptotically independent:

$$\mathcal{L}(X_1^N(t), \dots, X_k^N(t)) \longrightarrow \mu_t^{\otimes k}.$$

It goes back to Boltzmann, in order to close some equations.

Dobrushin 70'. Assume that the drift is Lipschitz in

$$W_1(\mu_s^N, \mu_s) \leq C_T \, W_1(\mu_0^N, \mu_0)$$

It is a coupling argument. To simplify notations, the write $W(x-y) = b(x,y)$ for the drift.

Denote $(Y_t^1, \dots, Y_t^N) \in (\mathbb{R}^d)^N$ defined as the McKean-Vlasov solution to

$$\mathrm{d}Y_t^i = b(Y_t^i \mathrm{d}t, \mu_t) + \mathrm{d}B_t^i, \quad \mathcal{L}(Y_t) = \mu_t^{\otimes N},$$

and

$$\mathrm{d}X_i(t) = -b(X_i(t), \mu_t^N)\mathrm{d}t + \mathrm{d}B_t^i$$

with $X_0 = Y_0$, and the same brownian motions. Then, a computation shows that

$$\begin{aligned}
\mathrm{d}\left(X_i(t) - Y_i(t)\right) &= -\frac{1}{N}\sum_{j=1}^{n} b(X_i(t), X_j(t)) - \int b(Y_i(t), y)\mathrm{d}\mu_t(\mathrm{d}y) \\
&= \frac{1}{N}\sum_{j=1}^{N} b(X_i(t), X_j(t)) - b(Y_i(t), X_j(t)) + \frac{1}{N}\sum_{i=1}^{N} b(Y_i(t), X_j(t)) - b(Y_i(t), Y_j(\\
&+ \frac{1}{N}\sum_{j=1}^{N} (b(Y_i(t), Y_j(t)) - \int b(Y_i^t, y)\mu_t(\mathrm{d}y)
\end{aligned}$$

We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|X_i(t) - Y_i(t)\| \leq \|b\|_{lip} \left(\|X_i(t) - Y_i(t)\| + \frac{1}{N}\sum_{j=1}^{N} \|X_j(t) - Y_j(t)\| \right)$$
$$+ |\frac{1}{N}\sum_{j=1}^{N}(b(Y_i(t), Y_j(t)) - \int b(Y_i^t, y)\mu_t(\mathrm{d}y)|$$

We conclude by a Grönwall, and a kind of Law of large number for the last term.

Question

Can we understand long time behavior?

In fact this PDE is a **Wasserstein Gradient Flow** as introduced by Jordan, Kinderlehrer, Otto, 1996' and Otto 01'

$$\partial_t \mu_t + \operatorname{div}(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)) = 0,$$

where $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ is called a **free energy**

Complicated non-linear PDE $\iff$ Optimization on measures

Think measures as an ensemble of particles as in Hydrodynamics, then the evolution of these particles satisfies **conservation of mass** i.e.

$$\partial \mu_t + \mathrm{div}(\mu_t v_t) = 0,$$

with v_t a velocity field. Among all possibles curves, particles minimize kinetic energy.

Benamou–Brenier dynamic formulation

$$W_2^2(\rho_0, \rho_1) = \inf_{(\rho_t, v_t)} \int_0^1 \int |v_t(x)|^2 \rho_t(\mathrm{d}x)\mathrm{d}t,$$

subject to

$$\partial_t \rho_t + \mathrm{div}(\rho_t v_t) = 0, \qquad \rho_{t=0} = \rho_0, \qquad \rho_{t=1} = \rho_1.$$

$\rightarrow$ Kind of Geodesic structure ? But What is the Tangent space ??

1st Try: Since $\mathcal{P}(\mathbb{R}^d)$ is convex, we obtain all the Radon measures with integral 0.
2nd Try: Let select the element that minimize the Benamou Brenier action, it gives

$$T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) = \{\nabla\psi | \psi : \mathbb{R}^d \to \mathbb{R} \text{ compact support, smooth}\}^{\bar{L^2(\mu)}},$$

and with Onsager metric

$$\langle \nabla\psi_1, \nabla\psi_2 \rangle_\mu = \int \langle \nabla\psi_1, \nabla\psi_2 \rangle \mathrm{d}\mu.$$

For a functional $\mathcal{F}$ on probability measures,

$$\partial_t \mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu_t) \mathrm{div}(\mu_t v_t) = -\int \langle \nabla\delta\mathcal{F}(\mu_t), v_t \rangle \mu_t = \langle \nabla\delta\mathcal{F}(\mu_t), v_t \rangle_{\mu_t}$$

the Wasserstein gradient is represented by

$$\nabla_{W_2}\mathcal{F}(\mu) = -\nabla \frac{\delta\mathcal{F}}{\delta\mu}.$$

The McKean–Vlasov PDE is a gradient flow

Define the free energy

$$\mathcal{F}_\beta(\mu) = \int V \,\mathrm{d}\mu + \frac{1}{2}\iint W(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \beta^{-1}\int \rho\log\rho\,\mathrm{d}x, \qquad \mu = \rho\mathrm{d}x.$$

Then

$$\frac{\delta\mathcal{F}_\beta}{\delta\mu} = V + W * \mu + \beta^{-1}\log\rho,$$

therefore

$$\partial_t\mu_t = \operatorname{div}\left(\mu_t\nabla\frac{\delta\mathcal{F}_\beta}{\delta\mu}(\mu_t)\right) = \operatorname{div}\big(\mu_t(\nabla V + \nabla W * \mu_t)\big) + \beta^{-1}\Delta\mu_t.$$

Long-time behavior: the useful energy identity

Along smooth solutions,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\beta(\mu_t) = -\int \left|\nabla \frac{\delta \mathcal{F}_\beta}{\delta \mu}(\mu_t)\right|^2 \mathrm{d}\mu_t \leq 0.$$

If a functional inequality holds, for instance a free-energy PL inequality

$$\int \left|\nabla \frac{\delta \mathcal{F}_\beta}{\delta \mu}\right|^2 \mathrm{d}\mu \geq 2\lambda\big(\mathcal{F}_\beta(\mu) - \mathcal{F}_\beta(\mu_*)\big),$$

then

$$\mathcal{F}_\beta(\mu_t) - \mathcal{F}_\beta(\mu_*) \leq e^{-2\lambda t}\big(\mathcal{F}_\beta(\mu_0) - \mathcal{F}_\beta(\mu_*)\big).$$

This inequality is called **LSI** for diffusions (See A. Surin for further details)

Common noise: the new modeling choice

Now assume particles share the same environmental Brownian motion B^0:

$$\mathrm{d}X_i(t) = b(X_i(t), \mu_t^N)\mathrm{d}t + \sigma(X_i(t), \mu_t^N)\mathrm{d}B_t^0.$$

Examples where this is natural:

- mean-field games with common;
- stochastic flows in random environments;
- homogenized deep models where the same random layer acts on all tokens.

The key difference is not just technical: all particles see the same randomness, so averaging over particles does not average out the noise.

The empirical measure now has a stochastic limit

For a smooth test function φ, Ito's formula gives

$$\mathrm{d}\left\langle \mu_t^N, \varphi \right\rangle = \left\langle \mu_t^N, b(\cdot, \mu_t^N) \cdot \nabla\varphi + \frac{1}{2}(\sigma\sigma^\top)(\cdot, \mu_t^N) : D^2\varphi, \mathrm{d} \right\rangle t$$
$$+ \left\langle \mu_t^N, \sigma(\cdot, \mu_t^N)^\top \nabla\varphi, \mathrm{d} \right\rangle B_t^0.$$

There is no $N^{-1/2}$ in front of the last term. The limit is a measure-valued SPDE.

The limiting nonlinear process is usually written as

$$\mathrm{d}X_t = b(X_t, \mu_t)\mathrm{d}t + \sigma(X_t, \mu_t)\mathrm{d}B_t^0, \qquad \mu_t = \mathcal{L}(X_t \mid B^0).$$

Then, for all smooth φ,

$$\begin{aligned}\mathrm{d}\left\langle \mu_t, \varphi \right\rangle = &\left\langle \mu_t, b(\cdot, \mu_t) \cdot \nabla\varphi + \frac{1}{2}(\sigma\sigma^\top)(\cdot, \mu_t) : D^2\varphi, \mathrm{d} \right\rangle t \\ &+ \left\langle \mu_t, \sigma(\cdot, \mu_t)^\top \nabla\varphi, \mathrm{d} \right\rangle B_t^0.\end{aligned}$$

The law of one particle is deterministic only after averaging over the environment. The conditional law is random.

Connection with homogenized Transformer limits

In a deep random architecture, all tokens in a layer share the same random weights. In a diffusion limit, this naturally produces common noise.

A schematic geometric limit on the sphere has the form

$$\mathrm{d}x_i(t) = P_{x_i(t)} B_i(X(t))\mathrm{d}t + \sum_{\alpha} P_{x_i(t)} \sigma_{\alpha,i}(X(t)) \circ \mathrm{d}W_t^{\alpha}, \qquad P_x = \mathrm{Id} - xx^{\top}.$$

Taking $N \to \infty$ then gives a random measure dynamics, not just a deterministic transport equation.

This is why conditional laws, stochastic flows, and SPDE methods become central.

Toy Model Transformers

Define $\kappa(t) := \frac{\alpha}{d}\|\frac{1}{n}\sum_{i=1}^{n} x_i(t)\|^2$, the dynamics is given by

$$dx_i(t) = \sqrt{\kappa(t)}P_{x_i(t)}dB_t - \frac{\kappa(t)}{2}(d-1)x_i(t)dt,$$

where B_t is a Brownian motion in $\mathbb{R}^d$.

Toy Model Transformers

Define $\kappa(t) := \frac{\alpha}{d} \| \frac{1}{n} \sum_{i=1}^{n} x_i(t) \|^2$, the dynamics is given by

$$dx_i(t) = \sqrt{\kappa(t)} P_{x_i(t)} dB_t - \frac{\kappa(t)}{2}(d-1)x_i(t)dt,$$

where B_t is a Brownian motion in $\mathbb{R}^d$.

- Interpretation: The tokens evolve as a (shared) Brownian motion on the sphere (up to a time change).

Toy Model Transformers

Define $\kappa(t) := \frac{\alpha}{d}\|\frac{1}{n}\sum_{i=1}^{n} x_i(t)\|^2$, the dynamics is given by

$$dx_i(t) = \sqrt{\kappa(t)}P_{x_i(t)}dB_t - \frac{\kappa(t)}{2}(d-1)x_i(t)dt,$$

where B_t is a Brownian motion in $\mathbb{R}^d$.

- Interpretation: The tokens evolve as a (shared) Brownian motion on the sphere (up to a time change).
- The law of X_i is uniform on the sphere.

Let look at the law of $(x_i(t), x_j(t))$, and only at $R_{ij}(t)\langle x_i(t), x_j(t)\rangle$

The angles evolve as

$$dR_{i,j}(t) = d\,\kappa(t)\,\left(1 - R_{i,j}(t) + \frac{R_{i,j}^2 + R_{i,j} - 2}{d}\right)$$
$$+ \sqrt{\kappa(t)}\left(\langle P_{x_i(t)} dB_t, x_j(t)\rangle + \langle P_{x_j(t)} dB_t, x_i(t)\rangle\right)$$

Let look at the law of $(x_i(t), x_j(t))$, and only at $R_{ij}(t)\langle x_i(t), x_j(t)\rangle$

The angles evolve as

$$dR_{i,j}(t) = d\,\kappa(t)\,\left(1 - R_{i,j}(t) + \frac{R_{i,j}^2 + R_{i,j} - 2}{d}\right) + \sqrt{\kappa(t)}\left(\langle P_{x_i(t)} dB_t, x_j(t)\rangle + \langle P_{x_j(t)} dB_t, x_i(t)\rangle\right)$$

$\rightarrow$ As $d \rightarrow +\infty$,

$$dR_{ij}(t) \sim d\kappa(t)(1 - R_{i,j}(t)) + \sqrt{\kappa(t)}\left(\langle P_{x_i(t)} dB_t, x_j(t)\rangle + \langle P_{x_j(t)} dB_t, x_i(t)\rangle\right)$$

Let look at the law of $(x_i(t), x_j(t))$, and only at $R_{ij}(t)\langle x_i(t), x_j(t)\rangle$

The angles evolve as

$$dR_{i,j}(t) = d\,\kappa(t)\,\left(1 - R_{i,j}(t) + \frac{R_{i,j}^2 + R_{i,j} - 2}{d}\right) + \sqrt{\kappa(t)}\left(\langle P_{x_i(t)} dB_t, x_j(t)\rangle + \langle P_{x_j(t)} dB_t, x_i(t)\rangle\right)$$

$\rightarrow$ As $d \to +\infty$,

$$dR_{ij}(t) \sim d\kappa(t)(1 - R_{i,j}(t)) + \sqrt{\kappa(t)}\left(\langle P_{x_i(t)} dB_t, x_j(t)\rangle + \langle P_{x_j(t)} dB_t, x_i(t)\rangle\right)$$

$\rightarrow$ As $d \to +\infty$, the martingale part is vanishing then

$$dR_{ij}(t) \sim d\kappa(t)(1 - R_{i,j}(t))$$

Common noise can change qualitative behavior

In common-noise systems, the same randomness can instead correlate particles. A stochastic flow may contract distances between two particles, producing synchronization by noise.

A typical phenomenon is:

one-point motion is ergodic, two-point motion contracts: $d(X_t, Y_t) \to 0$.

Noise-induced uniqueness of equilibrium

Gess–Gvalani–Martini 26' study the McKean–Vlasov SPDE on $\mathbb{T}^d$, $d \geq 2$,

$$\partial_t \rho = \Delta \rho + \nabla \cdot (\rho(\nabla W * \rho)) + \sqrt{2K}\, \nabla \cdot (\rho \circ \xi^\theta),$$

where ξ^θ is a common, white-in-time, spatially coloured, divergence-free transport noise.

Main theorem, informal If the noise is sufficiently mixing and the intensity K is large enough, then

the only invariant probability measure is δ_1.

Here 1 denotes the uniform density on $\mathbb{T}^d$.

Without the common noise, the deterministic McKean–Vlasov PDE can have several steady states because of phase transitions. The common transport noise destroys these competing equilibria at the level of the random dynamics.

What replaces the free-energy picture?

If the common noise is thermal

- possible random / stochastic gradient-flow formulation;
- random energy landscape for each environment;
- hope for pathwise energy identities.

If the noise is not thermal

- no reason for Gibbs equilibrium;
- invariant object may be a random measure;
- synchronization and clustering can dominate.

Thank you

Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you
Thank you Thank you Thank you Thank you Thank you Thank you Thank you Thank you