

GRADIENT FLOWS IN DATA SCIENCE

(Franca Hoffmann, California Institut of Technology)

LECTURE 1

Particle Methods for Bayesian Inverse Problems

AGENDA

1. Interacting Particle Methods in Data Science
2. Overview of Topics
3. Introduction to Bayesian Inverse Problems
4. Ensemble Kalman Inversion
5. Ensemble Kalman Sampler

Key References Lecture 1:

- Course by Amit Apte
- "Inverse problems: A Bayesian perspective" by A.M. Stuart (2010)
- "Inverse Problems and Data Assimilation" by Sanz-Alonso, Stuart, Taeb (2018)
- "Interacting Langerin Diffusions: Gradient Structure and Ensemble Kalman Sampler"
Garbuno-Inigo, Hoffmann, Li, Stuart (2020)
- "Ensemble Kalman Inversion: Mean-field Limit and convergence analysis"
Zhiyan Ding, Qin Li (2021)
- "Ensemble Kalman Sampler: Mean-field Limit and convergence analysis"
Zhiyan Ding, Qin Li (2021)
- "Sharp Propagation of Chaos for the Ensemble Langerin Sampler"
Urbain Vaes (2024)

1. Interacting Particle Methods in Data Science

Many data science tasks can be formulated as optimization problems or sampling problems

We will also consider filtering problems.

Algorithms aim to find approximate solutions, considering trade-offs between accuracy and computational cost \ speed.

Often, we may want to combine models with data, but our models may not fully represent the real world, and our data may be sparse and/or noisy.

Optimization Consider an objective function $J: \mathbb{R}^d \rightarrow \mathbb{R}$.

GOAL: Find x^* such that $J(x^*) = \min_{x \in \mathbb{R}^d} J(x)$.

CHALLENGES: J may be

- non-convex
- only given as black-box oracle
- not differentiable
- subject to constraints
- high-dimensional ($d \gg 1$)

← Such problems are pervasive in many fields, such as machine learning, signal/image processing, optimal control, weather forecast, climate analysis, engineering design, ...

IDEA: use a swarm of particles $\{x_i(t)\}_{i=1}^N \in \mathbb{R}^d$ that approximate x^* dynamically as $t \rightarrow T$ or $t \rightarrow \infty$.

Examples :

- Gradient-based methods (SGD, Adam, RMSProp, ...)
- Metaheuristics (often gradient-free)

- ⊕ Lots of freedom in design: exploration vs exploitation, use special structure of J , design adapted to available properties & information
- ⊖ Despite significant empirical success, most results are experimental in nature and lack rigorous mathematical foundation.

Sampling Consider a target measure $\mu \in \mathcal{P}(\mathbb{R}^d)$

GOAL: Create samples $\{x_i\}_{i=1}^N \sim \mu$.

→ can be used to approximate quantities of interest, e.g. $\mathbb{E}_\mu[f(x)] \sim \frac{1}{N} \sum_{i=1}^N f(x_i)$

CHALLENGE: This is a hard problem in general!

- If d large, we need a lot of particles for a good approximation:

$$\mathbb{E} W_2\left(\frac{1}{N} \sum_{i=1}^N \delta_{x_i}, \mu\right) \sim N^{-1/d} \quad [\text{Fournier - Guillin 2015}]$$

→ to half the error, one needs 2^d more samples

- μ is often only known up to normalization

→ to compute $\int \mu$, one needs to approximate the integral

- Rare high-probability regions go unseen, e.g. if the typical set (where μ has most of its mass) is far away from a narrow high-density mode.

→ many algorithms work well if $\mu \approx$ Gaussian, but do poorly away from that setting

- μ has structure at multiple length scales

- even if μ is nice, e.g. $\mu \ll \mathbb{Z}^d$ with density $e^{-J(x)}$, the potential J may have any of the challenges listed above

IDEA: Use a swarm of particles $\{x_i(t)\}_{i=1}^N$ for which $x_i(0)$ are easy to obtain (e.g. sampled from a Gaussian, or chosen deterministically) and $x_i(t)$ become approximate samples of μ as $t \rightarrow T$ or $t \rightarrow \infty$.

Example (Gaussian) Sampling from a Gaussian is easy!

Given $\mu = \mathcal{N}(m, C)$, let A such that $AA^T = C$.

If $z_i \sim \mathcal{N}(0, 1)$ for $z_i \in \mathbb{R}$, then $z = (z_1, \dots, z_d) \sim \mathcal{N}(0, \text{Id})$ and $x = m + Az \sim \mathcal{N}(m, C)$

To obtain z_i use inverse CDF: sample $u_i \sim \text{Unif}(0, 1)$, then $z_i = \Phi^{-1}(u_i) \sim \mathcal{N}(0, 1)$, where $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ Gaussian CDF. can use pseudorandom generator

$z_i := \Phi^{-1}(u_i) \sim \mathcal{N}(0, 1)$, where $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ Gaussian CDF.

↑ many algorithms exist to approximate this efficiently

Note: Choice of A is non-unique. Popular choices are $A =$ lower triangular (Cholesky factorization) or $A = C^{1/2}$ (symmetric square root)

Example) (Overdamped Langevin dynamics) Consider $\mu = e^{-J(x)} \in \mathcal{P}(\mathbb{R}^d)$.

Fix $\{x_i(0)\}_{i=1}^N \sim \rho_0$ iid, and define $x_i(t)$ as solutions to

$$dx_i(t) = -\nabla J(x_i(t))dt + \sqrt{2} dB_i(t) \quad (1)$$

where $B_i(t)$ is a d -dimensional Brownian motion.

Define the empirical measure $\rho^N(t) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}$.

Note that $x_i(t)$ are iid $\forall t \geq 0$ as there is no coupling between particles.

Let $\rho_t = \text{Law}(x_i(t))$. Then ρ_t solves the Fokker-Planck equation

$$\partial_t \rho = \text{div}(\rho \nabla J + \nabla \rho) = \text{div}(\rho \nabla \log \frac{\rho}{\mu}) \quad (2)$$

The unique steady state of (2) is $\rho = \mu$, and $\rho_t \rightarrow \mu$ in W_1 as $t \rightarrow \infty$ if J has sufficient convexity.

Mean-field limit: $\rho^N(t) \rightarrow \rho(t)$ as $N \rightarrow \infty$

Advantages:

- Easy to implement
- Well-developed mathematical theory exists

Draw-backs:

- Need ∇J evaluations to run particles

- Need convexity of J for convergence: $D^2 J \geq \lambda \text{Id} > 0$

- Need ∇J Lipschitz for well-posedness of (1).

← or $J = \text{convex} + \text{bounded}$
← or dissipativity (convex outside a ball)

→ many alternatives to Overdamped Langevin exist

→ important to understand this toy sampler to compare with more advanced methods

Take-away: Interacting particle systems are a powerful tool for designing optimization and sampling algorithms!

⇒ PDE analysis useful

Remarks

Another example of interacting particle systems for sampling: Generative Modelling

- Here, the choice of the dynamics is entirely up to us.

No physical constraints! Of course, we can take inspiration from nature to achieve certain desirable properties of the algorithm.

- **Highly parallelizable!**

- Theoretical guarantees can often be derived via **mean-field descriptions**.

Add explanation errors: $\Delta t \rightarrow 0, N \rightarrow \infty, t \rightarrow \infty$ via triangle ineq.

2. Overview of Topics

L1 (Monday) : Particle Methods for Bayesian Inverse Problems

- Introduction to Bayesian Inverse Problems
- Ensemble Kalman Inversion
- Ensemble Kalman Sampler

L2 (Tuesday) : Covariance-modulated optimal transport

L3 (Wednesday)

- Introduction and Definition
- Application to Bayesian Inverse Problems
- Formal W_c gradient flow structure
- Splitting in shape and moments
- Variance-modulated OT
- Existence of W_c geodesics
- Dual Formulation of W_c
- Convergence to Equilibrium in W_c gradient flows

L4 (Thursday) : Propagation of Chaos, Large Time Behaviour & Accuracy for Filtering Algorithms

L5 (Friday)

- The Ensemble Kalman Filter in the near Gaussian setting
- Propagation of Chaos for Transport Ensemble Filters
- Large-time behaviour for the Mean-Field Ensemble Kalman Filter

L6 (Friday) : Consensus-Based Optimization and Sampling

- Intro to Metaheuristic algorithms and CBX
- Large-time behaviour
- Uniform-in-time Propagation of Chaos
- Variants and relation to particle swarm optimization

3. Introduction to Bayesian Inverse Problems

Consider the inverse problem

$$y = G(x) + \eta$$

- $x \in \mathbb{R}^d$ parameter (unknown)
- $y \in \mathbb{R}^k$ observations / measurements (given)
- $G: \mathbb{R}^d \rightarrow \mathbb{R}^k$ forward operator (given)
- η observation noise (distribution known)

Prior information on x : $x \sim \mathcal{N}(m_0, \Sigma_0)$

Observation noise: $\eta \sim \mathcal{N}(0, T)$, $T \in \mathbb{R}^{k \times k}$, $T \succ 0$

Ref: [Stuart 2010], [Dashti, Stuart 2017]

$\Rightarrow$ By Baye's Theorem, the posterior distribution $\Pi(x|y)$ is given by

$$\Pi(x) = \Pi(x|y) \propto e^{-J(x)}, \quad J(x) = \underbrace{\frac{1}{2} \|y - G(x)\|_T^2}_{\text{log likelihood}} + \underbrace{\frac{1}{2} \|x - m_0\|_{\Sigma_0}^2}_{\text{prior}}$$

When solving Bayesian inverse problems, one has typically one of two goals:

(I) Inversion: find $x^* = \arg\max_{x \in \mathbb{R}^d} \Pi(x)$ "MAP estimator" (maximizing $\Pi(x|y)$)
 $\rightarrow$ most likely x under model and data

(II) Sampling from $\Pi \rightarrow$ understand the shape of $\Pi(x|y)$
 $\rightarrow$ approximate quantities of interest $\mathbb{E}_\Pi[f]$.

Note that (I) is an optimization problem

and (II) is at the core of uncertainty quantification (UQ).

Example 1 The forward map G is linear: $G(x) = Hx$, $H \in \mathbb{R}^{d \times k}$

Then $J(x) = \frac{1}{2} \|y - Hx\|_T^2 + \frac{1}{2} \|x - m_0\|_{\Sigma_0}^2$ is quadratic

$\Rightarrow \Pi$ is Gaussian: $\mathcal{N}(m, C)$ with $m = m_0 + K(y - Hm_0)$
 $C = (I - KH)\Sigma_0$

where $K = \Sigma_0 H^T (H \Sigma_0 H^T + T)^{-1}$ is the Kalman gain

$\Rightarrow$ Bayesian IP is explicitly solvable

Premark: If G is not differentiable, or DG is too costly to obtain, we need derivative-free methods to solve (1) and (2).

$\Rightarrow$ Ensemble Kalman Methods are popular algorithms in Bayesian Inversion.

We will introduce two recent methods:

- Ensemble Kalman Inversion (EKI) $\Rightarrow$ solves (I)
- Ensemble Kalman Sampler (EKS) $\Rightarrow$ solves (II)

4. Ensemble Kalman Inversion

Ensemble Kalman Inversion (EKI) is a derivative-free optimization algorithm to invert G :

$$dx_i(t) = -\frac{1}{N} \sum_{k=1}^N \underbrace{\langle G(x_k(t)) - \bar{G}, G(x_i(t)) - y \rangle_T}_{\text{drive particles to match the data}} x_k(t) dt + \underbrace{C^{xG}(X) T^{-1} \sqrt{\Sigma}}_{\substack{\text{choice} \\ \in \mathbb{R}^k}} dB_i(t) \quad (1)$$

[Schillings, Stuart 2017]

where $\bar{x}(t) = \frac{1}{N} \sum_{k=1}^N x_k(t)$, $\bar{G}(t) = \frac{1}{N} \sum_{k=1}^N G(x_k(t))$, $\Sigma \in \mathbb{R}^{k \times k}$ can be chosen freely, and

$$C^{xG}(X) = \frac{1}{N} \sum_{k=1}^N (x_k(t) - \bar{x}(t)) (G(x_k(t)) - \bar{G}(t))^T \in \mathbb{R}^{d \times k}$$

Properties:

- $\Sigma = 0$: optimization algorithm
- $\Sigma = T$ and $G(x) = Hx$ is linear: sampling algorithm mapping $x_i(0) \sim \text{prior}$ into $x_i(1) \sim \text{posterior}$ [Chen, Oliver 2012]
 - ⊕ reach posterior samples in finite time (not at $t \rightarrow \infty$ as for Langevin)
 - ⊖ need to initialize with samples from the prior
- $\Sigma = T$: as $t \rightarrow \infty$, solutions collapse to a single point solving the opt. pb (I)

Linear Case

let prior $= N(m_0, \Sigma_0)$, $G(x) = Hx$ linear, $\Sigma = \Gamma$ in EKI.

Then posterior $= N(m_1, C_1)$ with m_1, C_1 as in Example 1:

$$\begin{cases} m_1 = m_0 + K(y - Hm_0), \text{ with Kalman gain } K = \Sigma_0 H^T (H \Sigma_0 H^T + \Gamma)^{-1} \\ C_1 = (I - KH) \Sigma_0, \end{cases}$$

In this case, the mean $m(t)$ and covariance $C(t)$ satisfy closed equations

$$\begin{cases} \dot{m} = -C(H^T \Gamma^{-1} H m - r), \quad r = H^T \Gamma^{-1} y \\ \dot{C} = -C H^T \Gamma^{-1} H C \end{cases} \quad \left. \vphantom{\begin{cases} \dot{m} \\ \dot{C} \end{cases}} \right\} m(t), C(t) \text{ are mean-field objects}$$

• Solving for the covariance $C(t)$:

The inverse covariance satisfies $\frac{d}{dt} \bar{C} = -\bar{C} \dot{C} \bar{C} = H^T \Gamma^{-1} H$

$\Rightarrow C(t)^{-1}$ grows linearly in time

If $C(0) = \Sigma_0$, then $\bar{C}(t) = \Sigma_0^{-1} + H^T \Gamma^{-1} H t$, and so $C(t)$ is the posterior covariance

As $t \rightarrow \infty$, $C(t) \rightarrow 0$ (i.e. particles collapse to consensus)

• Solving for the mean $m(t)$:

$$\frac{d}{dt} (\bar{C} m) = -\bar{C} \dot{C} \bar{C} m + \bar{C} \dot{m} = H^T \Gamma^{-1} H m - H^T \Gamma^{-1} H m + r = r$$

$$\begin{aligned} \text{If } m(0) = m_0, \text{ then } \bar{C}_t m_t &= \Sigma_0^{-1} m_0 + r t \Rightarrow m(t) = (\Sigma_0^{-1} + H^T \Gamma^{-1} H t)^{-1} (\Sigma_0^{-1} m_0 + r t) \\ \Rightarrow m(1) &= (\Sigma_0^{-1} + H^T \Gamma^{-1} H)^{-1} (\Sigma_0^{-1} m_0 + H^T \Gamma^{-1} y) = (\Sigma_0^{-1} + H^T \Gamma^{-1} H t)^{-1} (\Sigma_0^{-1} m_0 + H^T \Gamma^{-1} y t) \\ &= m_0 + (\Sigma_0^{-1} + H^T \Gamma^{-1} H)^{-1} H^T \Gamma^{-1} (y - H m_0) \\ &= m_0 + K(y - H m_0) \quad \text{since } (\Sigma_0^{-1} + H^T \Gamma^{-1} H)^{-1} H^T \Gamma^{-1} = \Sigma_0 H^T (H \Sigma_0 H^T + \Gamma)^{-1} \end{aligned}$$

$$\Rightarrow t \rightarrow \infty: m(t) = (\frac{1}{t} \Sigma_0^{-1} + H^T \Gamma^{-1} H)^{-1} (\frac{1}{t} \Sigma_0^{-1} m_0 + H^T \Gamma^{-1} y)$$

if $H^T \Gamma^{-1} H$ is invertible, then $m(t) \rightarrow (H^T \Gamma^{-1} H)^{-1} H^T \Gamma^{-1} y$ as $t \rightarrow \infty$

least squares optimizer / MLE (prior disappears)

else, let $V_0 = \text{Ker } H^T \Gamma^{-1} H$ and decompose $\mathbb{R}^d = V_0 \oplus V_0^\perp$. Then $m(t) = m(t)|_{V_0^\perp} + m(t)|_{V_0}$,

$$m(t)|_{V_0^\perp} \rightarrow \arg \min_{x \in V_0^\perp} \frac{1}{2} \|Hx - y\|_{\Gamma}^2 = (H^T \Gamma^{-1} H)^\dagger H^T \Gamma^{-1} y =: m_\infty^\perp \quad \text{Moore-Penrose Pseudoinverse}$$

$$m(t)|_{V_0} \rightarrow m_0|_{V_0} + (P \Sigma_0^{-1} P)^{-1} P \Sigma_0^{-1} Q (m_0|_{V_0^\perp} - m_\infty^\perp) \quad \text{where } Px = x|_{V_0}, Qx = x|_{V_0^\perp}$$

If $P \Sigma_0^{-1} Q = 0$ (eg. $\Sigma_0 = \sigma^2 I$)

then
 $m(t)|_{V_0} = m_0|_{V_0}$
 prior mean on
 unidentifiable
 subspace

on directions where data provides no information ($H^T \Gamma^{-1} y|_{V_0} = 0$), Σ_0^{-1} never gets dominated, no matter how large t is.

$\Rightarrow$ identifiable ("data-informed") directions converge to their MLE values

unidentifiable ("ill-posed") directions retain prior information

This is a general principle in inverse problems: for rank-deficient (ill-posed) forward maps, some prior regularization is retained, even in the infinite-data limit!

Mean-Field Limit of EK1

Theorem 1 | Assume $G(x) = Hx + u(x)$, where $u: \mathbb{R}^d \rightarrow \mathbb{R}^k$ smooth bounded

Let $\mu^N(t) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}$ for $x_i(t)$ solving (1) with $\Sigma = T$ and $x_i(0) \sim \rho_0$.

Then $\mu^N(t) \rightarrow \rho(t)$ solving the PDE

$$\partial_t \rho = -\operatorname{div}(\rho(y - G(x))^T T^{-1} \operatorname{Cov}_{G,\rho}) + \frac{1}{2} \operatorname{Tr}[\operatorname{Cov}_{\rho,G} T^{-1} \operatorname{Cov}_{G,\rho} D^2 \rho]$$

Assume $\rho_0 \in C^2$ and $\forall p > 0, \int |x|^p d\rho_0(x) < \infty$.

Then $\forall \varepsilon > 0 \exists C_\varepsilon(t)$ independent of N s.t.

$$\mathbb{E}[W_2(\mu^N(t), \rho(t))] \leq C_\varepsilon(t) \begin{cases} N^{-\frac{1}{2}+\varepsilon} & \text{if } d \leq 4 \\ N^{-2/d} & \text{if } d > 4 \end{cases}$$

Here,

$$\operatorname{Cov}_{G,\rho} = \int (G(x) - \mathbb{E}_\rho[G(X)])(x - \mathbb{E}_\rho[X])^T d\rho(x) \in \mathbb{R}^{k \times d}$$

$$\operatorname{Cov}_{\rho,G} = \int (x - \mathbb{E}_\rho[X])(G(x) - \mathbb{E}_\rho[G(X)])^T d\rho(x) \in \mathbb{R}^{d \times k}$$

Proof: See [Ding, Li 2021]

Remark • get $N^{-\frac{1}{2}+\varepsilon}$ rate for any $d \geq 1$ if we look at weak convergence
in duality with Lipschitz functions
• holds for slightly non-Gaussian initial data as well

Open Problems:

Results for more nonlinear $G(x)$

→ mean-field limit (quantitative rates, uniform-in-time)

→ how close is $\rho(t, x)$ to $\pi(x)$, and $\lim_{t \rightarrow \infty} \rho(t, x)$ to $x^* = \operatorname{argmin} J(x)$?

5. Ensemble Kalman Sampler

Ensemble Kalman Sampler (EKS) is a derivative-free sampling algorithm that approximates samples from the Bayesian posterior distribution $\pi(x) = e^{-J(x)}$.

IDEA: Start with EKI and turn it into a sampler for $t \rightarrow \infty$ by adding noise to prevent the particle collapse

$$dx_i(t) = \underbrace{-\frac{1}{N} \sum_{k=1}^N \langle G(x_k(t)) - \bar{G}, G(x_i(t)) - y \rangle_T}_{\text{drive particles to consensus}} x_k(t) dt + \underbrace{C(p^N(t)) \sum_{k=1}^N (x_k(t) - m_0)}_{\text{drive particles to match the data}} dt + \underbrace{\sqrt{2C(p^N(t))}}_{\text{regularization from prior}} dB_i(t) \quad (2)$$

to prevent particle collapse

where $C(p^N) = \int (x - \mathbb{E}_{p^N}[X]) (x - \mathbb{E}_{p^N}[X])^T dp^N(x) = \frac{1}{N} \sum_{k=1}^N (x_k - \bar{x})(x_k - \bar{x})^T$

Here, noise is added in the parameter space, rather than the data space.

(Benefits of this choice were demonstrated in the context of optimization in [Kovachki, Stuart 2019].)

Relation of EKS to preconditioned gradient descent:

Assume

$$G(x_i) - \bar{G} \approx H(x_i - \bar{x}), \text{ where } H = DG(\bar{x})$$

Substitute into first term in (2):

Justified if

- $|x_i - x_j|$ small $\forall i, j$
- $\{x_i\}$ concentrated in region where $DG(x) \approx \text{constant}$

$$\begin{aligned} -\frac{1}{N} \sum_{k=1}^N \langle G(x_k) - \bar{G}, G(x_i) - y \rangle_T x_k &= -\frac{1}{N} \sum_{k=1}^N \langle G(x_k) - \bar{G}, G(x_i) - y \rangle_T (x_k - \bar{x}) \\ &\approx -\frac{1}{N} \sum_{k=1}^N \langle H(x_k - \bar{x}), Hx_i - y \rangle_T (x_k - \bar{x}) = -\frac{1}{N} \sum_{k=1}^N \langle x_k - \bar{x}, H^T T^{-1} (Hx_i - y) \rangle (x_k - \bar{x}) \\ &= -\frac{1}{N} \sum_{k=1}^N [(x_k - \bar{x}) \otimes (x_k - \bar{x})] H^T T^{-1} (Hx_i - y) = -C(p^N) \nabla \frac{1}{2} \|Hx_i - y\|_T^2 \end{aligned}$$

Since $\nabla \frac{1}{2} \|x - m_0\|_{\Sigma_0}^2 = \Sigma_0^{-1} (x - m_0)$, EKS simplifies to

$$dx_i(t) = -C(p^n(t)) \nabla \tilde{J}(x_i(t)) dt + \sqrt{2C(p^n(t))} dB_i(t) \quad (3)$$

where $\tilde{J}(x) = \frac{1}{2} \|Hx - y\|_T^2 + \frac{1}{2} \|x - m_0\|_{\Sigma_0}^2$ is a quadratic approximation of $J(x)$ by linearizing J

- Remark:
- If G is linear, (3) is exactly EKS
 - If G is nonlinear, but differentiable, we can still consider (3) with J instead of $\tilde{J} \rightarrow$ alternative algorithm:
 dynamically pre-conditioned Langevin MCMC method
 ($C(p^n)$ acts as a surrogate for $D^2 J(x)^{-1}$)
 - Can consider (2) a way of approximating ∇J with differences

Linear Case Let $G(x) = Hx$. Then $\Pi(x) = N(m_\infty, C_\infty)$ with $(m_\infty, C_\infty) = (m, C)$ from Example 1.

EKS reduces to (3) with $\tilde{J}(x) = J(x) = \frac{1}{2} \|x - m_\infty\|_{C_\infty}^2$. Then the mean and covariance solve
 mean-field objects

$$\begin{cases} \dot{m}(t) = -C(t) (C_\infty^{-1} m - r) & \text{with } r = C_\infty^{-1} m_\infty \\ \dot{C}(t) = -2C(t) C_\infty^{-1} C(t) + 2C(t) & m(0) = m_0, \quad C(0) = \Sigma_0 \end{cases}$$

Note that $\frac{d}{dt} C(t)^{-1} = -C^{-1} \dot{C} C^{-1} = 2C_\infty^{-1} - 2\bar{C}^{-1}(t) \Rightarrow C(t)^{-1} = (\Sigma_0^{-1} - C_\infty^{-1}) e^{-2t} + C_\infty^{-1}$

$$\begin{aligned} \text{Then } \frac{d}{dt} \|m(t) - m_\infty\|_{C(t)}^2 &= 2 \langle \dot{m}, \bar{C}^{-1}(m - m_\infty) \rangle - \langle m - m_\infty, \bar{C}^{-1} \dot{C} \bar{C}^{-1} (m - m_\infty) \rangle \\ &= -2 \langle C_\infty^{-1} m - r, (m - m_\infty) \rangle + 2 \langle m - m_\infty, C_\infty^{-1} (m - m_\infty) \rangle - 2 \langle m - m_\infty, \bar{C}^{-1} (m - m_\infty) \rangle \\ &= -2 \|m(t) - m_\infty\|_{C(t)}^2 \Rightarrow \|m(t) - m_\infty\|_{C(t)} = \|m(0) - m_\infty\|_{\Sigma_0} e^{-t} \end{aligned}$$

Hence as $t \rightarrow \infty$: $m(t) \rightarrow m_\infty$, $C(t) \rightarrow C_\infty$ exponentially fast

Mean-Field Limit of EKS

Consider the SDE

$$dx_i(t) = -C(p^n(t)) \nabla J(x_i(t)) dt + \sqrt{2C(p^n(t))} dB_i(t) \quad (4)$$

and its mean-field analogue

$$\begin{cases} d\bar{x}(t) = -C(p(t)) \nabla J(\bar{x}(t)) dt + \sqrt{2C(p(t))} dB(t) \\ p(t) = \text{Law}(\bar{x}(t)) \end{cases} \quad (5)$$

C is a quadratic quantity: cannot directly apply
 Skorohod's lemma

The law $p(t)$ solves the nonlinear Fokker-Planck equation

$$\partial_t p = \operatorname{div}(p C(p) \nabla J + C(p) \nabla p) \quad (6)$$

Theorem 2 | Assume $J \in C^2(\mathbb{R}^d)$ and for some $\ell \geq 0$

$$\left. \begin{aligned} & \bullet J(x) \sim |x|^{\ell+2} \\ & \bullet |\nabla J(x)| \sim |x|^{\ell+1} \\ & \bullet c_- |x|^\ell \operatorname{Id} \leq D^2 J(x) \leq c_+ |x|^\ell \operatorname{Id} \end{aligned} \right\} \begin{aligned} & \text{for large enough } |x| \\ & \text{and } c_\pm > 0 \end{aligned}$$

Assume $\rho_0 \in \mathcal{P}_q(\mathbb{R}^d) \forall q \geq 1$ and $C(\rho_0) > 0$. Then $\forall p > 0 \exists C > 0$ s.t.

$$\mathbb{E} \left[\sup_{t \in [0, T]} |x_i(t) - \bar{x}_i(t)|^p \right] \leq C N^{-1/2}$$

pathwise estimate for W_p
[Vaes 2024]

Remarks • Assumptions hold in particular for $\rho_0 = N(m_0, \Sigma_0)$.

• Since $C(p)$ is a quadratic quantity in drift and diffusion,

one cannot directly apply Snitzmann's synchronous coupling approach

→ For $G(x) = Hx$, [Ding, Li (2021)] employed a novel bootstrapping argument from [Chaintron, Diez] and obtained

$$\sup_{t \in [0, T]} \left(\mathbb{E} |x_i(t) - \bar{x}_i(t)|^2 \right)^{1/2} \leq C J^{-1/2 + \varepsilon} \quad \forall \varepsilon > 0$$

pointwise estimate for W_2

→ [Vaes (2024)] obtained the sharp rate

using stopping times, which originates in numerical analysis [Higham, Mao, Stuart (2002)]
novel for application in mean-field limits

• Moment assumptions on ρ_0 can be relaxed.

• Wasserstein- p empirical chaos follows from the triangle inequality:

$$\mathbb{E}[W_p(\rho^n(t), \rho(t))] \leq \underbrace{\mathbb{E}[W_p(\rho^n(t), \bar{\rho}^n(t))]}_{\text{controlled by Thm 2}} + \underbrace{\mathbb{E}[W_p(\bar{\rho}^n(t), \rho(t))]}_{\leq J^{-\alpha} \text{ for some } \alpha = \alpha(p, d, q) \in (0, \frac{1}{2}] \text{ if } \int |x|^q \rho_t(x) dx < \infty \text{ [Fournier, Guillin (2015)]}} \lesssim J^{-\alpha}$$

$$\text{where } \rho^n(t) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}, \bar{\rho}^n(t) = \frac{1}{N} \sum_{i=1}^N \delta_{\bar{x}_i(t)}$$

→ Overall rate dominated by sampling rate.

• Both Theorem 1 (EKI) and Theorem 2 (EKS) are **finite-time results**:

$$C = C_T \rightarrow \infty \text{ as } T \rightarrow \infty$$

Open question: Uniform-in-time propagation of chaos