Sampling Via Gradient Flows In The Space of Probability Measures

(With Links To Interacting Particle Systems)

Andrew Stuart, Caltech

Topics on Neuroscience, Collective Migration and Parameter Estimation

Oxford, July 5^{th} 2023.

Collaborators

Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance

https://arxiv.org/abs/2302.11024 [9]

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, Andrew M. Stuart

Review paper: Trillos, Hosseini, Sanz-Alonso [30] (2023)

Outline

Outline

Goal

The Sampling Problem

$V : \mathbb{R}^d \to \mathbb{R}$. Draw (approximate) samples from

$$\rho^\star(\theta) \propto \exp\Big(-V(\theta)\Big)$$

MCMC: Brooks, Galin, Jones, Meng [6] (2011)

SMC: Del Moral, Doucet, Jasra [10] (2006)

Unifying Framework

Ingredients For Gradient Flows

- $L^2 = L^2(\mathbb{R}^d; \mathbb{R})$
- $\mathcal{P} = L^2$ Probability Densities on $\mathbb{R}^d$
- $\mathcal{E} : \mathcal{P} \to \mathbb{R}^+,\ \mathcal{E}(\rho^\star) = 0$ (Energy Functional)
- $\frac{\delta \mathcal{E}}{\delta \rho} \in L^2$ (First Variation)
- $M(\rho) : L^2 \to L^2$ invertible, positive semi-definite for all $\rho \in \mathcal{P}$

Nonlinearly Preconditioned Gradient Flow in $\mathcal{P}$

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)$$

Key Identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\rho_t) = \left\langle \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t), \frac{\partial \rho_t}{\partial t} \right\rangle_{L^2} = -\left\langle M(\rho_t)\frac{\partial \rho_t}{\partial t}, \frac{\partial \rho_t}{\partial t} \right\rangle_{L^2} \leq 0$$

Gradient Flows: Ambrosio, Gigili, Savaré [3] (2005).

Sampling via optimization: Wibisono [33] (2018).

Canonical Example 1

At Our Disposal: Energy Functional $\mathcal{E}(\cdot)$, Metric $M(\cdot)$.

Energy Functional

Kullback–Leibler (KL) Divergence $\mathcal{E} : \mathcal{P} \to \mathbb{R}^+$, $\mathcal{E}(\rho^\star) = 0$, $\rho^\star = \operatorname{argmin}_{\rho \in \mathcal{P}} \mathcal{E}(\rho)$:

$$\mathcal{E}(\rho) = \mathrm{KL}[\rho \| \rho^\star] = \int \rho \log\left(\frac{\rho}{\rho^\star}\right) \mathrm{d}\theta$$

$$\frac{\delta \mathcal{E}}{\delta \rho}(\rho; \rho^\star) = \log \rho - \log \rho^\star + \text{constant}$$

Metric

Wasserstein−2 Metric Tensor:

$$M(\rho)^{-1}\psi = -\nabla_\theta \cdot (\rho \nabla_\theta \psi)$$

Canonical Example 2

Gradient Flow: Fokker-Planck Equation

KL for energy: $\mathcal{E} = \mathrm{KL}[\rho \| \rho^\star]$; Wasserstein$-2$ for metric; then:

$$\frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t \nabla_\theta \log \rho^\star) + \nabla_\theta \cdot (\rho_t \nabla_\theta \log \rho_t)$$
$$\frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t \nabla_\theta \log \rho^\star) + \nabla_\theta \cdot (\nabla_\theta \rho_t)$$

Trivial Mean Field Model: Langevin Equation

$\mathrm{Law}(\theta_t) = \rho_t$:

$$\mathrm{d}\theta_t = \nabla_\theta \log \rho^\star(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck and Langevin equations: Risken [26] (1996), Pavliotis [23] (2014)

Fokker-Planck as gradient flow for $\mathcal{E}(\rho)$: Jordan, Kinderlehrer, Otto [15] (1998)

Langevin equation and MCMC: Roberts, Tweedie [28] (1996); Roberts, Rosenthal [27] (2001)

Canonical Example 3

Theorem Markowich, Villani [21] (2000);

Assume $\exists \lambda > 0$:

$$D^2 V(\cdot) \succeq \lambda I$$

Then, for all $t \geq 0$,

$$\mathrm{KL}[\rho_t \| \rho^\star] \leq \mathrm{KL}[\rho_0 \| \rho^\star] e^{-2\lambda t}$$

Rate of exponential convergence depends on problem

Probabilistic methods: Mattingly, S and Higham [22] (2002)

Outline

Choice of $\mathcal{E}$

f-divergence

Consider f: $f(1) = 0$ and f convex and define:

$$D_f[\rho \| \rho^\star] = \int \rho^\star f\left(\frac{\rho}{\rho^\star}\right) \mathrm{d}\theta$$

Examples

- Kullback–Leibler divergence: $f(x) = x \log x$
- χ^2 divergence: $f(x) = (x-1)^2$
- Hellinger distance: $f(x) = (\sqrt{x} - 1)^2$
- ...

Choice of $\mathcal{E}$: Kullback–Leibler (KL) is Special

Energy: Kullback-Leibler

$$\mathcal{E}(\rho;\rho^\star) = \mathrm{KL}[\rho\|\rho^\star] = \int \rho \log\left(\frac{\rho}{\rho^\star}\right) \mathrm{d}\theta$$

$$\frac{\delta\mathcal{E}}{\delta\rho}(\rho;\rho^\star) = \log\rho - \log\rho^\star + \text{constant}$$

$$\mathcal{E}(\rho;c\rho^\star) = \mathcal{E}(\rho;\rho^\star) - \log(c)$$

Theorem Chen, Huang, Huang, Reich, AMS [9] (2023)

KL is the only f-divergence whose first variation leads to a gradient flow which is independent of the normalization constant of ρ^*

Use Kullback-Leibler from now on

Outline

Two Metrics

Wasserstein Metric Jordan, Kindelehrer, Otto [15] (1998)

$$\text{Metric: } M(\rho)^{-1}\psi = -\nabla_\theta \cdot (\rho \nabla_\theta \psi)$$

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t \nabla_\theta \log \rho^\star) + \nabla_\theta \cdot (\nabla_\theta \rho_t)$$

$$\text{Trivial Mean Field Model: } \mathrm{d}\theta_t = \nabla_\theta \log \rho^\star(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fisher-Rao Metric Rao [24] (1945); Amari [1] (1998)

$$\text{Metric: } M(\rho)^{-1}\psi = \rho(\psi - \mathbb{E}_\rho[\psi])$$

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho^\star - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho^\star - \log \rho_t]$$

Nontrivial Mean Field Models: discuss later

Optimal transport: Villani [31] (2008)

Information geometry: Amari [2] (2016); Ay, Jost, Lê, Schwachhöfer [4] (2017)

Fisher-Rao Flow: Invariance Under Diffeomorphisms

Pushforward

Given diffeomorphism $\varphi : \mathbb{R}^d \to \mathbb{R}^d$

- $\tilde{\rho}_t = \varphi_{\#}\rho_t$ is the transformed distribution at time t
- $\tilde{\rho}^\star = \varphi_{\#}\rho^\star$ is the transformed target distribution

Proposition

Fisher-Rao gradient flow is invariant under any diffeomorphism:

$$\frac{\partial \rho_t}{\partial t} = \rho_t\left(\log \rho^\star - \log \rho_t\right) - \rho_t \mathbb{E}_{\rho_t}[\log \rho^\star - \log \rho_t]$$

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \tilde{\rho}_t\left(\log \tilde{\rho}^\star - \log \tilde{\rho}_t\right) - \tilde{\rho}_t \mathbb{E}_{\tilde{\rho}_t}[\log \tilde{\rho}^\star - \log \tilde{\rho}_t]$$

Consequence of Invariance of Fisher-Rao Gradient Flow

Theorem Lu, Slepčev, Wang [19] (2022); Chen, Huang, Huang, Reich, AMS [9] (2023)

Assume

- $\exists K > 0$:

$$e^{-K(1+|\theta|^2)} \leq \frac{\rho_0(\theta)}{\rho^\star(\theta)} \leq e^{K(1+|\theta|^2)}$$

- $\exists B > 0$ bounding first and second moments of $\rho_0, \rho^\star$

Then, for all $t \geq \log\big((1+B)K\big)$,

$$\mathrm{KL}[\rho_t \| \rho^\star] \leq (2 + B + eB)Ke^{-t}$$

Unconditional uniform exponential convergence

Mean-Field Models For Fisher-Rao Gradient Flow

Mean-Field ODE Chen, Huang, Huang, Reich, AMS [9] (2023)

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\nabla_\theta F(\theta; \rho_t, \rho^\star)\big|_{\theta=\theta_t}$$

$$-\nabla_\theta \cdot \Big(\rho(\theta)\nabla_\theta F(\theta; \rho, \rho^\star)\Big) = \rho(\theta)\mathbb{E}_\rho\Big(\log \rho^\star - \log \rho\Big) - \rho(\theta)\Big(\log \rho^\star(\theta) - \log \rho(\theta)\Big)$$

Particle approximation: $\{\theta_{t,\ell}\}_{\ell=1}^N$

Birth-Death Process Lu, Lu, Nolen [18] (2019); Lu, Slepčev, Wang [19] (2022)

$$\Omega_t^\ell = \log\Big(\frac{1}{N}\sum_{j=1}^N K(\theta_{t,\ell} - \theta_{t,j})/\rho^\star(\theta_{t,\ell})\Big), \quad K \approx \delta$$

$$\Lambda_t^i = \Omega_t^i - \frac{1}{N}\sum_{\ell=1}^N \Omega_t^\ell \quad \text{Particle } i \text{ birth-death rate}$$

Both face significant obstacles in order to implement

Outline

Invariance Revisited

Theorem Ay, Jost, Lê, Schwachhöfer [4] (2015); Bauer, Bruveris, Michor [5] (2016); Cencov [8] (2000)

The Fisher-Rao metric is the only Riemannian metric on smooth positive densities (up to scaling) that is invariant under any diffeomorphism of the parameter space

Affine Invariance

Given an affine transformation $\varphi : \mathbb{R}^d \to \mathbb{R}^d$

- $\tilde{\rho}_t = \varphi_{\#}\rho_t$ is the transformed distribution at time t
- $\tilde{\rho}^\star = \varphi_{\#}\rho^\star$ is the transformed target distribution

Flow is affine invariant if, for all affine φ, $(\tilde{\rho}_t, \tilde{\rho}^\star)$ satisfy same equation as $(\rho_t, \rho^\star)$.

For parallel MCMC: Goodman, Weare [13] (2010); generalization: Leimkuhler, Matthews, Weare [17] (2018)

For ensemble Kalman: Garbuno-Inigo, Nüsken and Reich [12] (2020)

For ensemble Kalman: Huang, Huang, Reich, AMS [14] (2022)

Examples

Fisher-Rao Gradient Flow

The Fisher-Rao gradient flow is affine invariant

Kalman-Wasserstein Gradient Flow Garbuno-Inigo, Hoffman, Li and AMS [11] (2020)

The Kalman-Wasserstein gradient flow is affine invariant.

$$\begin{aligned}
\text{Covariance: } & C(\rho) = \mathrm{Cov}(\rho) \\
\text{Metric: } & M(\rho)^{-1}\psi = -\nabla_\theta \cdot (\rho C(\rho) \nabla_\theta \psi) \\
\text{Flow: } & \frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t C(\rho_t) \nabla_\theta \log \rho^\star) + \nabla_\theta \cdot (C(\rho_t) \nabla_\theta \rho_t) \\
\text{Mean Field Model: } & \mathrm{d}\theta_t = C(\rho_t) \nabla_\theta \log \rho^\star(\theta_t) \mathrm{d}t + \sqrt{2C(\rho_t)} \mathrm{d}W_t
\end{aligned}$$

Kalman-Wasserstein metric first identified: Reich and Cotter [25] (2015)

Consequence of Affine Invariance of Kalman-Wasserstein Gradient Flow

Theorem Garbuno-Inigo, Hoffman, Li and AMS [11] (2022); Carrillo and Vaes [7] (2023)

Assume V is quadratic. Then there is constant $C > 0$ such that, for all $t \geq 0$,

$$\mathcal{W}_2(\rho_t, \rho^\star) \leq C\mathcal{W}_2(\rho_0, \rho^\star)e^{-t}$$

Unconditional uniform exponential convergence

Universal convergence to equilibirum for Gaussian targets: Garbuno-Inigo, Hoffman, Li and AMS [11] (2020)

Universal convergence to equilibirum for Gaussian targets (non-Gaussian initialization): Carrillo and Vaes [7] (2021)

Numerical Example Illustrating Affine Invariance

Experimental Set-Up

- 2D Rosenbrock potential:

$$V(\theta) = \frac{\lambda}{20}\left(\theta_2 - \theta_1^2\right)^2 + \frac{1}{20}\left(1 - \theta_1\right)^2$$

  for $\theta = (\theta_1, \theta_2)$ and $\lambda = 10^{-k}$, $k = 0, 1, 2$

- Goal: sample $\rho^\star \propto \exp(-V(\theta))$
- Method 1: Wasserstein using noninteracting Langvein, 10^3 particles.
- Method 2: Kalman-Wasserstein using interacting Langevin, 10^3 particles
- Configuration: Integrate to $t = 15$, initialized from

$$\theta_0 \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}\right)$$

Numerical Example Illustrating Affine Invariance

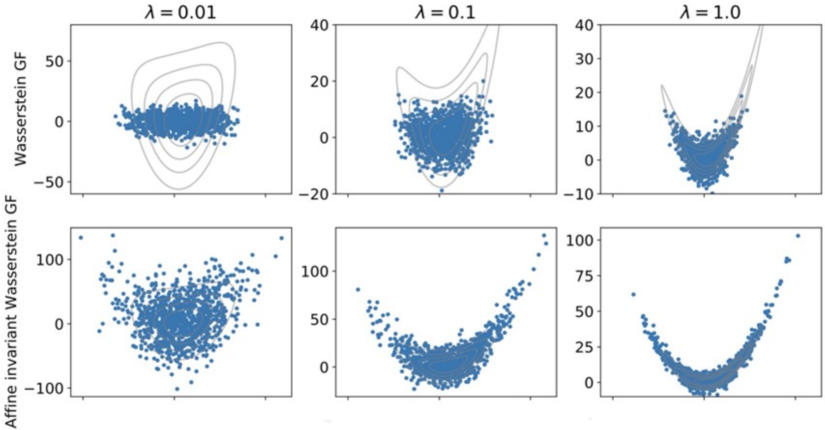

Figure: 10^3 particles at $t = 15$ from Langevin (top row) and affine invariant Langevin (bottom row). Grey lines represent the contour of the true posterior

Outline

Variational Bayes

Energy Functional

Kullback–Leibler (KL) Divergence:

$$\mathcal{E}(\rho) = \mathrm{KL}[\rho \| \rho^\star] = \int \rho \log\left(\frac{\rho}{\rho^\star}\right) \mathrm{d}\theta$$

- $\mathcal{E} : \mathcal{P} \to \mathbb{R}^+, \; \mathcal{E}(\rho^\star) = 0$
- $\rho^\star = \operatorname{argmin}_{\rho \in \mathcal{P}} \mathcal{E}(\rho)$
- $\mathcal{P}_a$: parameterized subset of probability density functions on $\mathbb{R}^d$, $a \in \mathbb{R}^p$
- $\rho_{a_\star} = \operatorname{argmin}_{\rho \in \mathcal{P}_a} \mathcal{E}(\rho)$

Variational Bayes: Mackay [20] (2008); Wainright, Jordan [32] (2008)

Gradient Descent for Variational Bayes

Ingredients For Gradient Flows

- $\mathcal{E} : \mathcal{P} \to \mathbb{R}_+, \ \mathcal{E}(\rho^\star) = 0$ (Energy Functional)
- $\mathcal{P}_a \subset \mathcal{P}, \ a \in \mathbb{R}^p, \ \rho(a) \in \mathcal{P}_a$ (Candidate Density)
- $\langle M(\rho)\nabla_a \rho(a) \cdot \sigma_1, \nabla_a \rho(a) \cdot \sigma_2 \rangle_{L^2} = \langle \mathfrak{M}(a)\sigma_1, \sigma_2 \rangle_{\mathbb{R}^p}$ (Induced Metric)

The Gradient Flow in $\mathbb{R}^p$ (and in $\mathcal{P}_a$)

$$\frac{\mathrm{d}}{\mathrm{d}t} a_t = -\mathfrak{M}(a_t)^{-1} \frac{\partial}{\partial a} \mathcal{E}(\rho_a)\Big|_{a=a_t}$$

Identifying The Gradient Flow: Gaussian Case 1

Example: Gaussian Variational Bayes

- $\mathcal{G}$: all Gaussian probability measures on $\mathbb{R}^d$
- $\mathcal{G} = \mathcal{P}_a$, $a = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}_{\mathrm{sym},\geq 0}$

Theorem Chen, Huang, Huang, Reich, AMS [9] (2023)

Moment closure gives the gradient flow

Identifying The Gradient Flow: Gaussian Case 2

Consequence

- Consider a gradient flow in $\mathcal{P}$:

$$\frac{\partial \rho_t(\theta)}{\partial t} = \sigma_t(\theta, \rho_t)$$

.

- Then mean and covariance evolve according to

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = \int \sigma_t(\theta, \rho_t)\theta \mathrm{d}\theta, \qquad \frac{\mathrm{d}C_t}{\mathrm{d}t} = \int \sigma_t(\theta, \rho_t)(\theta - m_t)(\theta - m_t)^T \mathrm{d}\theta$$

- Closure: to obtain gradient flow in $\mathcal{P}_a$ use $\rho_t = \rho_{a_t} = \mathcal{N}(m_t, C_t)$

Moment closure in variational Kalman filtering: Särkkä [29] (2007)

Moment closure in Wasserstein gradient flow: Lambert, Chewi, Bach, Bonnabel, Rigollet [16] (2022)

Convergence Rates

Theorem Chen, Huang, Huang, Reich, AMS [9] (2023)

Assume Gaussian target $\rho^\star$ and consider Fisher-Rao variational inference. If $\rho^\star = \mathcal{N}(m_\star, C_\star)$, and $C_0 = \lambda_0 I, \lambda_0 > 0$, then

$$\|m_t - m_\star\|_2 = \Theta(e^{-t}), \quad \|C_t - C_\star\|_2 = \Theta(e^{-t})$$

See also: Lambert, Chewi, Bach, Bonnabel, Rigollet [16] (2022)

Numerical Example: Gaussian Gradient Flows

Experimental Set-Up

- 2D convex potential:

$$V(\theta) = \frac{1}{20}(\sqrt{\lambda}\theta_1 - \theta_2)^2 + \frac{1}{20}(\theta_2)^4$$

  for $\theta = (\theta_1, \theta_2)$ and $\lambda = 10^{-k}$, $k = 0, 1, 2$

- Goal: sample $\rho^\star \propto \exp(-V(\theta))$
- Method 1: Gaussian approximation of Fisher-Rao GF
- Method 2: Gaussian approximation of Wasserstein GF
- Method 3: Gaussian approximation of vanilla GF
- Configuration: Integrate to $t = 15$ initialized from the Gaussian

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}\right)$$

Numerical Examples

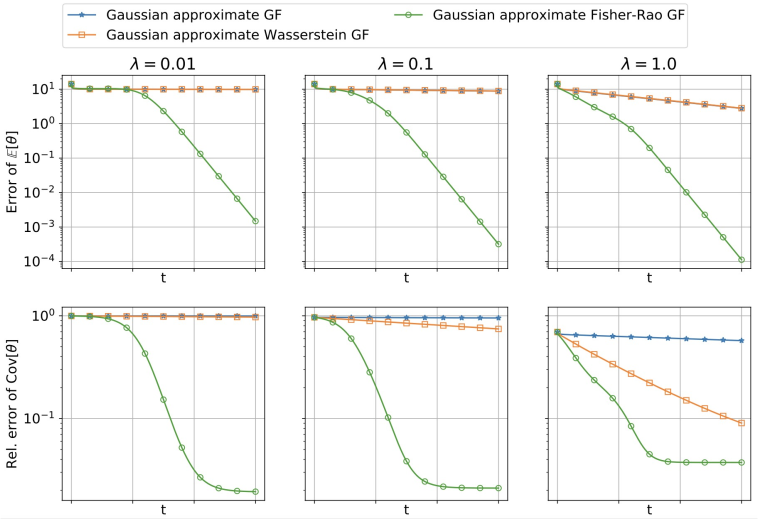

Figure: x axis is from $t = 0$ to 15. Gaussian approximate Fisher-Rao gradient flows perform the best. Convergence rate not influenced by different values of λ

Outline

Summary

Gradient Flows for Sampling Chen, Huang, Huang, Reich, AMS [9] (2023)

- Energy Functional: KL divergence
  - invariant to normalization consts
  - unique property among all f divergences
- Fisher-Rao Metric:
  - invariant to any diffeomorphism of the parameters
  - unique property among all metrics on probability space
  - uniform exponential convergence
  - implementing mean field models is difficult
  - works well within Gaussian variational inference
- Affine Invariance:
  - uniform exponential convergence for Gaussian target
  - affine invariant Kalman-Wasserstein
  - implementation of mean field models is straightforward

Thank-you

https://arxiv.org/abs/2302.11024 [9]

Gradient flows for sampling: mean-field models, Gaussian approximations and affine invariance

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, Andrew M. Stuart

References I

[1] S.-I. Amari.
Natural gradient works efficiently in learning.
Neural computation, 10(2):251–276, 1998.

[2] S.-i. Amari.
Information Geometry and its Applications, volume 194.
Springer, 2016.

[3] L. Ambrosio, N. Gigli, and G. Savaré.
Gradient flows: in Metric Spaces and in the Space of Probability Measures.
Springer Science & Business Media, 2005.

[4] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer.
Information Geometry, volume 64.
Springer, 2017.

[5] M. Bauer, M. Bruveris, and P. W. Michor.
Uniqueness of the Fisher–Rao metric on the space of smooth densities.
Bulletin of the London Mathematical Society, 48(3):499–506, 2016.

[6] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng.
Handbook of Markov chain Monte Carlo.
CRC press, 2011.

References II

[7] J. Carrillo and U. Vaes.
Wasserstein stability estimates for covariance-preconditioned fokker–planck equations.
Nonlinearity, 34(4):2275, 2021.

[8] N. N. Cencov.
Statistical decision rules and optimal inference.
American Mathematical Soc., 2000.

[9] Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart.
Gradient flows for sampling: Mean-field models, gaussian approximations and affine invariance.
arXiv preprint arXiv:2302.11024, 2023.

[10] P. Del Moral, A. Doucet, and A. Jasra.
Sequential Monte Carlo samplers.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):411–436, 2006.

[11] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart.
Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler.
SIAM Journal on Applied Dynamical Systems, 19(1):412–441, 2020.

References III

[12] A. Garbuno-Inigo, N. Nüsken, and S. Reich.
Affine invariant interacting Langevin dynamics for Bayesian inference.
SIAM Journal on Applied Dynamical Systems, 19(3):1633–1658, 2020.

[13] J. Goodman and J. Weare.
Ensemble samplers with affine invariance.
Communications in applied mathematics and computational science, 5(1):65–80, 2010.

[14] D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart.
Efficient derivative-free Bayesian inference for large-scale inverse problems.
arXiv preprint arXiv:2204.04386, 2022.

[15] R. Jordan, D. Kinderlehrer, and F. Otto.
The variational formulation of the Fokker–Planck equation.
SIAM journal on mathematical analysis, 29(1):1–17, 1998.

[16] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet.
Variational inference via Wasserstein gradient flows.
arXiv preprint arXiv:2205.15902, 2022.

[17] B. Leimkuhler, C. Matthews, and J. Weare.
Ensemble preconditioning for Markov chain Monte Carlo simulations.
Stat. Comput., 28:277–290, 2018.

References IV

[18] J. Lu, Y. Lu, and J. Nolen.
Scaling limit of the Stein variational gradient descent: The mean field regime.
SIAM Journal on Mathematical Analysis, 51(2):648–671, 2019.

[19] Y. Lu, D. Slepčev, and L. Wang.
Birth-death dynamics for sampling: Global convergence, approximations and their asymptotics.
arXiv preprint arXiv:2211.00450, 2022.

[20] D. J. MacKay.
Information Theory, Inference and Learning Algorithms.
Cambridge University Press, 2003.

[21] P. A. Markowich and C. Villani.
On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis.
Mat. Contemp, 19:1–29, 2000.

[22] J. C. Mattingly, A. M. Stuart, and D. J. Higham.
Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise.
Stochastic processes and their applications, 101(2):185–232, 2002.

References V

[23] G. A. Pavliotis.
Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, volume 60.
Springer, 2014.

[24] C. R. Rao.
Information and the accuracy attainable in the estimation of statistical parameters.
Reson. J. Sci. Educ, 20:78–90, 1945.

[25] S. Reich and C. Cotter.
Probabilistic forecasting and Bayesian data assimilation.
Cambridge University Press, 2015.

[26] H. Risken.
Fokker-Planck Equation.
Springer, 1996.

[27] G. O. Roberts and J. S. Rosenthal.
Optimal scaling for various metropolis-hastings algorithms.
Statistical science, 16(4):351–367, 2001.

References VI

[28] G. O. Roberts and R. L. Tweedie.
Exponential convergence of Langevin distributions and their discrete approximations.
Bernoulli, pages 341–363, 1996.

[29] S. Särkkä.
On unscented Kalman filtering for state estimation of continuous-time nonlinear systems.
IEEE Transactions on automatic control, 52(9):1631–1641, 2007.

[30] N. G. Trillos, B. Hosseini, and D. Sanz-Alonso.
From optimization to sampling through gradient flows.
arXiv preprint arXiv:2302.11449 (To appear in Notices of AMS), 2023.

[31] C. Villani.
Optimal transport: old and new, volume 338.
Springer, 2009.

[32] M. J. Wainwright, M. I. Jordan, et al.
Graphical models, exponential families, and variational inference.
Foundations and Trends® in Machine Learning, 1(1–2):1–305, 2008.

References VII

[33] A. Wibisono.
Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.
In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.