Partial Differential Equations for sampling: A connection between Sampling and Gradient Flows

Francesca R. Crucinio ESOMAS, University of Turin & Collegio Carlo Alberto

Joint work with: Nicolas Chopin, Anna Korba, Sahani Pathiraja.

Introduction

Sampling

- **Aim:** sample from a probability distribution π on $\mathbb{R}^d$ and approximate expectations w.r.t. $\pi(x) = \gamma(x)/\mathcal{Z}$ whose normalising constant might be unknown

$$\int f(x)\pi(x)\mathrm{d}x$$

- **Motivation:** compute posterior expectations in Bayesian inference

When π is not available in closed form is natural to resort to numerical approximation methods.

Sampling as optimisation over distributions

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) := \int_{\mathbb{R}^d} \mu(x) \log \frac{\mu(x)}{\pi(x)} \mathrm{d}x$$

where $\mathrm{KL}(\mu|\pi)$ denotes the Kullback–Leibler divergence.

- Variational Inference
- Algorithms based on the Langevin diffusion: Random walk Metropolis (RWM), Metropolis adjusted Langevin algorithm (MALA), Unadjusted Langevin algorithm (ULA)
- Algorithms based on tempering: sequential Monte Carlo (SMC), Annealed importance sampling (AIS), Parallel tempering (PT)

Gradient Flows

Gradient descent in Euclidean space

Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^+$ be a functional on $\mathbb{R}^d$. Consider the optimisation problem

$$\min_{z \in \mathbb{R}^d} \mathcal{F}(z).$$

Gradient descent in Euclidean space

Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^+$ be a functional on $\mathbb{R}^d$. Consider the optimisation problem

$$\min_{z \in \mathbb{R}^d} \mathcal{F}(z).$$

The gradient descent ODE in **Euclidean space** is

$$\dot{x}_t = -\nabla \mathcal{F}(x_t).$$

An Euler discretisation of the above gives the standard gradient descent algorithm

$$x_{n+1} = x_n - \gamma_{n+1} \nabla \mathcal{F}(x_n).$$

Gradient descent on $\mathcal{P}(\mathbb{R}^d)$

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. Consider the optimisation problem

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi).$$

Gradient descent on $\mathcal{P}(\mathbb{R}^d)$

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. Consider the optimisation problem

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi).$$

Gradient descent in this space is given by the following gradient flow

$$\dot{\mu_t} = -\nabla_{\mathcal{M}} \mathrm{KL}(\mu_t|\pi)$$

where $\mathcal{M}$ denotes the metric w.r.t. which the gradient is taken.

Wasserstein Gradient Flow

If the metric is the Wasserstein-2 distance we obtain the **Wasserstein gradient flow PDE** (Jordan et al., 1998)

$$\begin{aligned}\partial_t \mu_t &= -\nabla_{W_2} \mathrm{KL}(\mu_t|\pi) \\ &= \mathrm{div}\left(\mu_t \nabla \log\left(\frac{\mu_t}{\pi}\right)\right) \\ &= -\mathrm{div}\left(\mu_t \nabla \log \pi\right) + \Delta \mu_t.\end{aligned}$$

Wasserstein Gradient Flow

If the metric is the Wasserstein-2 distance we obtain the **Wasserstein gradient flow PDE** (Jordan et al., 1998)

$$\begin{aligned}\partial_t \mu_t &= -\nabla_{W_2} \mathrm{KL}(\mu_t|\pi) \\ &= \mathrm{div}\left(\mu_t \nabla \log\left(\frac{\mu_t}{\pi}\right)\right) \\ &= -\mathrm{div}\left(\mu_t \nabla \log \pi\right) + \Delta \mu_t.\end{aligned}$$

Using the connection between Fokker–Plank PDEs and SDEs we obtain

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t$$

which is known as the **Langevin diffusion**.

Langevin based algorithms

Simple Euler–Maruyama discretisation leads to the **Unadjusted Langevin Algorithm** (ULA)

$$X_{n+1} = X_n + \gamma \nabla \log \pi(X_n) + \sqrt{2\gamma}\xi_{n+1}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

Langevin based algorithms

Simple Euler–Maruyama discretisation leads to the **Unadjusted Langevin Algorithm** (ULA)

$$X_{n+1} = X_n + \gamma \nabla \log \pi(X_n) + \sqrt{2\gamma}\xi_{n+1}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

Many others:

- Metropolis adjusted Langevin algorithm (MALA)
- Random walk Metropolis (RWM)

Fisher–Rao Gradient Flow

If the metric is the Fisher–Rao metric (or Hellinger) we obtain the **Fisher–Rao gradient flow PDE**

$$\partial_t \mu_t = \mu_t \left(\log\left(\frac{\pi}{\mu_t}\right) - \mathbb{E}_{\mu_t}\left[\log\left(\frac{\pi}{\mu_t}\right)\right]\right)$$

which has analytic solution (Chen et al., 2023)

$$\mu_t \propto \pi^{1-e^{-t}} \mu_0^{e^{-t}}.$$

Fisher–Rao Gradient Flow and Tempering

Consider a time discretisation of the Fisher–Rao gradient flow

$$\mu_n \propto \pi^{1-e^{-t_n}} \mu_0^{e^{-t_n}}.$$

Fisher–Rao Gradient Flow and Tempering

Consider a time discretisation of the Fisher–Rao gradient flow

$$\mu_n \propto \pi^{1-e^{-t_n}} \mu_0^{e^{-t_n}}.$$

In the Monte Carlo literature, it is common to consider the following **tempering (or annealing)** sequence

$$\mu_n \propto \mu_0^{1-\lambda_n} \pi^{\lambda_n},$$

where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = 1$. There are at the basis of

- Parallel Tempering
- Annealed Importance Sampling
- Sequential Monte Carlo tempering
- Termodynamic Integration

If $\lambda_n = 1 - e^{-t_n}$ for some t_n, tempering is a time discretisation of the Fisher–Rao gradient flow (Chopin et al., 2024; Domingo-Enrich and Pooladian, 2023).

Fisher–Rao with importance sampling

Time discretisation of the FR gradient flow: given μ_n we obtain μ_{n+1} as

$$\mu_{n+1}(x) \propto \pi(x)^{1-e^{-t_{n+1}}} \mu_0(x)^{e^{-t_{n+1}}} = \frac{\pi(x)^{1-e^{-t_{n+1}}}}{\pi(x)^{1-e^{-t_n}}} \frac{\mu_0(x)^{e^{-t_{n+1}}}}{\mu_0(x)^{e^{-t_n}}} \mu_0(x)^{e^{-t_n}} \pi(x)^{1-e^{-t_n}}$$

$$= \left(\frac{\pi(x)}{\mu_n(x)}\right)^{1-e^{-(t_n - t_{n+1})}} \mu_n(x) = \left(\frac{\pi(x)}{\mu_0(x)}\right)^{e^{-t_n} - e^{-t_{n+1}}} \mu_n(x).$$

If we have $X_n^1, \ldots, X_n^N \sim \mu_n$ we can approximate μ_{n+1} by importance sampling with weights

$$W_n^i = \left(\frac{\pi(X_n^i)}{\mu_0(X_n^i)}\right)^{e^{-t_n} - e^{-t_{n+1}}}$$

as in tempering SMC tempering/AIS.

Wasserstein or Fisher–Rao?

W If π satisfies a log-Sobolev inequality

$$\mathrm{KL}(\mu_t||\pi) \leq \mathrm{KL}(\mu_0||\pi)e^{-2\lambda_\pi^{-1}t}$$

FR If μ_0, π have bounded second moments and $|\log(\mu_0/\pi)| \leq M(1+|x|^2)$

$$\mathrm{KL}(\mu_t||\pi) \leq Ce^{-t}.$$

Solution: Wasserstein–Fisher–Rao gradient flow

Combining the W and FR dynamics we obtain the **WFR gradient flow**.

$$\partial_t \mu_t = \mu_t \left(\log\left(\frac{\pi}{\mu_t}\right) - \mathbb{E}_{\mu_t}\left[\log\left(\frac{\pi}{\mu_t}\right)\right] \right) + \nabla \cdot \left(\mu_t \nabla \log\left(\frac{\mu_t}{\pi}\right) \right).$$

Enjoys better convergence properties

$$\mathrm{KL}(\mu_t || \pi) \leq \min\left\{ \mathrm{KL}(\mu_t^{\mathrm{FR}} | \pi), \mathrm{KL}(\mu_t^{\mathrm{W}} | \pi) \right\}.$$

Sharper rates are work in progress (happy to chat about this!).

Wasserstein–Fisher–Rao gradient flow: Convergence

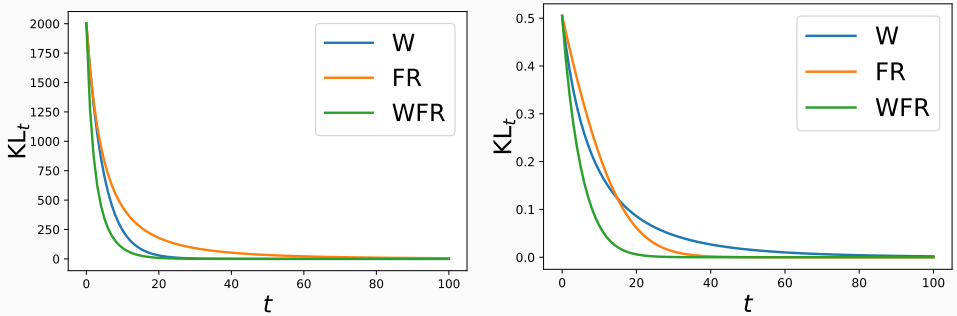


Figure 1: Evolution of KL along different PDE flows with $\mu_0(x) = \mathcal{N}(x; 0, 1)$ and $\pi(x) = \mathcal{N}(x; 20, 0.1)$ (left), $\pi(x) = \mathcal{N}(x; 1, 5)$ (right).

WFR with sequential Monte Carlo

A possible time discretisation of this PDE consists in applying the W flow first and then the FR flow. Given $X_0^1, \dots X_0^N \sim \mu_0$

1. propose new locations using ULA (W flow)

$$X_n^i = X_{n-1}^i + \gamma \nabla \log \pi(X_{n-1}^i) + \sqrt{2\gamma}\xi_n^i \qquad \xi_n^i \sim \mathcal{N}(0, \mathrm{Id}_d)$$

   and obtain $\mu_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_n^i}$

2. reweight (FR flow)

$$W_n^i = \left(\frac{\pi(X_n^i)}{\mu_n^N(X_n^i)}\right)^{1-e^{-(t_n - t_{n+1})}}.$$

WFR with sequential Monte Carlo

A possible time discretisation of this PDE consists in applying the W flow first and then the FR flow. Given $X_0^1, \dots X_0^N \sim \mu_0$

1. propose new locations using ULA (W flow)

$$X_n^i = X_{n-1}^i + \gamma \nabla \log \pi(X_{n-1}^i) + \sqrt{2\gamma}\xi_n^i \qquad \xi_n^i \sim \mathcal{N}(0, \mathrm{Id}_d)$$

   and obtain $\mu_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_n^i}$

2. reweight (FR flow)

$$W_n^i = \left(\frac{\pi(X_n^i)}{\mu_n^N(X_n^i)}\right)^{1-e^{-(t_n - t_{n+1})}}.$$

This algorithm falls within the class of **sequential Monte Carlo** methods

SMC-WFR

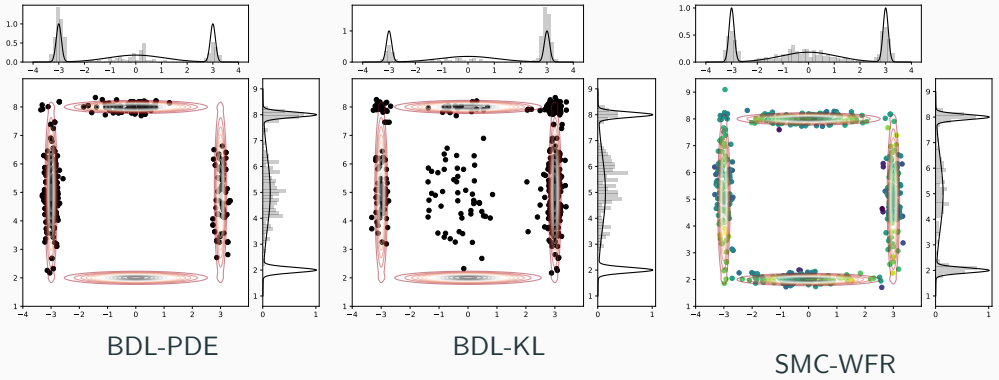


Figure 2: Comparison of approximations of the target $\pi(x) = \sum_{i=1}^{4} w_i \mathcal{N}(x; m_i, C_i)$ for the birth-death Langevin algorithms and our SMC approximation. For the latter the colour of the particles corresponds to the weight (brighter corresponds to higher weight). We compare both the joint distribution and the marginals.

Tempered Dynamics

Tempered Dynamics

Replace π with a time varying target $\pi_t \propto \pi^{\lambda_t} \mu_0^{1-\lambda_t}$ for $\lambda_t : \mathbb{R}_+ \to [0, 1]$

Tempered W

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \log \left(\frac{\mu_t}{\pi_t} \right) \right).$$

Can be implemented via Tempered Langevin dynamics.

Tempered FR

$$\partial_t \mu_t = \mu_t \left(\log \left(\frac{\pi_t}{\mu_t} \right) - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\pi_t}{\mu_t} \right) \right] \right).$$

Has analytic solution

$$\mu_t \propto \mu_0^{e^{-t} + \int_0^t e^{s-t}(1-\lambda_s) ds} \pi^{\int_0^t e^{s-t} \lambda_s ds}.$$

These are **not** gradient flows of $\mathrm{KL}(\mu|\pi)$.

Tempered Dynamics: Convergence

W

$$\mathrm{KL}(\mu_t||\pi) \leq \mathrm{KL}(\mu_0||\pi)e^{-2\lambda_\pi^{-1}t} + A\int_0^t e^{-2(t-s)\lambda_\pi^{-1}}(1-\lambda_s)ds$$

FR

$$\mathrm{KL}(\mu_t||\pi) \leq C(1-\int_0^t e^{s-t}\lambda_s ds).$$

Tempered Dynamics: Convergence

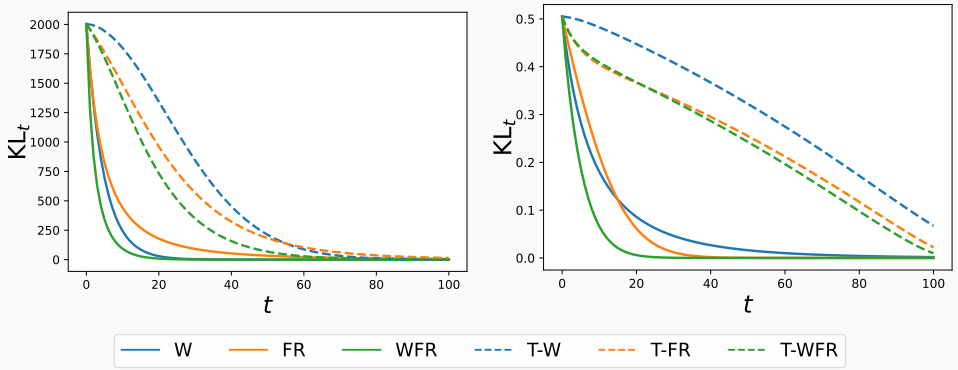


Figure 3: Evolution of KL along different PDE flows with $\mu_0(x) = \mathcal{N}(x; 0, 1)$ and $\pi(x) = \mathcal{N}(x; 20, 0.1)$ (left), $\pi(x) = \mathcal{N}(x; 1, 5)$ (right).

Conclusions

Metric	Dynamics	Algorithm
Wasserstein	Langevin diffusion	ULA (Markov chain)
Tempered Wasserstein	Tempered Langevin diffusion	Tempered ULA
Fisher–Rao	Tempering	Importance sampling
Tempered Fisher–Rao	Tempering	Importance sampling
WFR	Diffusion+Reweighting	sequential Monte Carlo

The gradient flow perspective is useful for

- deriving new algorithms (and justifying known ones!)
- convergence results
- opens the door to extensions through the use of other divergences

Splitting Wasserstein–Fisher–Rao

WFR PDE

The WFR PDE

$$\partial_t \mu_t = \mu_t \left(\log \left(\frac{\pi}{\mu_t} \right) - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\pi}{\mu_t} \right) \right] \right) + \nabla \cdot \left(\mu_t \nabla \log \left(\frac{\mu_t}{\pi} \right) \right)$$

naturally lends itself to splitting:

- the FR flow has known analytic solution $\mu_t \propto \pi^{1-e^{-t}} \mu_0^{e^{-t}}$ which can be approximated via IS
- the W flow can be approximated by e.g. ULA.

W-FR Splitting

A sequential splitting applied to the WFR PDE takes the form

$$\begin{aligned}
\hat{\nu}_1(x;\gamma) &= S_W(\gamma,\mu_0)(x)\\
\nu_1(x;\gamma) &= S_{FR}(\gamma,\hat{\nu}_1)(x)\\
\hat{\nu}_2(x;\gamma) &= S_W(\gamma,\nu_1)(x)\\
\nu_2(x;\gamma) &= S_{FR}(\gamma,\hat{\nu}_2)(x)\\
&\vdots
\end{aligned}$$

We denote the FR-W splitting by $\eta_i(x;\gamma)$.

S_W and S_{FR} do not commute, thus $\eta_i(x;\gamma) \neq \nu_i(x;\gamma)$ for $\gamma > 0$.

FR-W or W-FR?

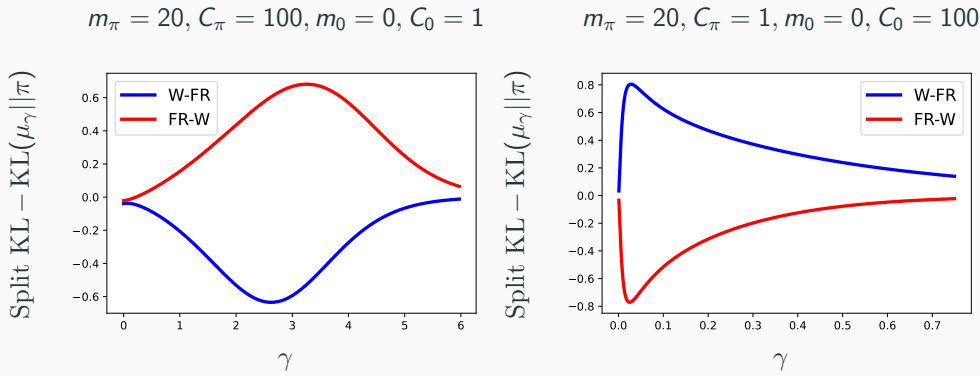


Figure 4: Difference in KL for a single time step γ for W-FR split and FR-W on 1D Gaussians. Left: Target more diffuse than initial distribution ($m_\pi = 20, C_\pi = 100, m_0 = 0, C_0 = 1$). Right: Target more concentrated than initial distribution ($m_\pi = 20, C_\pi = 1, m_0 = 0, C_0 = 100$).

Measuring Improvement

We aim at minimising $\mathrm{KL}(\mu|\pi)$. Check the time derivative over $[0, \gamma]$

WFR

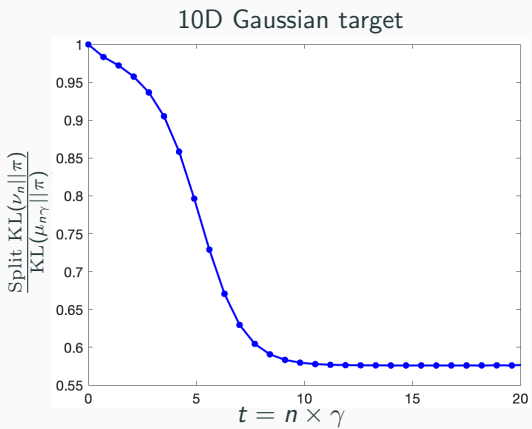
$$\frac{d}{d\gamma}\,\mathrm{KL}(\mu_\gamma|\pi) = -\mathbb{E}_{\mu_\gamma}\left[\left|\nabla \log \frac{\mu_\gamma}{\pi}\right|^2\right] - \mathrm{Var}_{\mu_\gamma}\left[\log \frac{\mu_\gamma}{\pi}\right]$$

W-FR

$$\frac{d}{d\gamma}\,\mathrm{KL}(\nu_\gamma|\pi) = -\mathbb{E}_{\nu_\gamma}\left[\left|\nabla \log \frac{\nu_\gamma}{\pi}\right|^2\right] - \mathrm{Var}_{\nu_\gamma}\left[\log \frac{\nu_\gamma}{\pi}\right] + (e^\gamma - 1)\mathrm{Cov}_{\nu_\gamma}\left(\log \frac{\nu_\gamma}{\pi}, \left|\nabla \log \frac{\nu_\gamma}{\pi}\right|^2\right)$$

FR-W

$$\begin{aligned}\frac{d}{d\gamma}\,\mathrm{KL}(\eta_\gamma||\pi) = &-\mathbb{E}_{\eta_\gamma}\left[\left|\nabla \log \frac{\eta_\gamma}{\pi}\right|^2\right] - \mathrm{Var}_{\eta_\gamma}\left[\log \frac{\eta_\gamma}{\pi}\right] \\ &- \int_0^\gamma \mathrm{Cov}_{\eta_\tau}\left(S_W\left(\gamma - \tau, \log \frac{\eta_\gamma}{\pi}\right), |\nabla \log \frac{\eta_\tau}{\pi}|^2\right) d\tau\end{aligned}$$



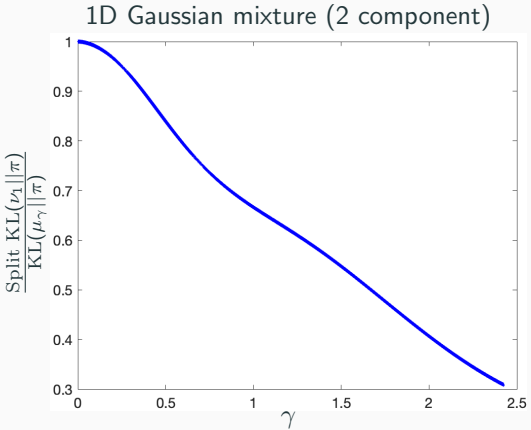


Figure 5: Left: Ratio of KL from n-step W-FR scheme to KL from continuous time WFR as a function of $t = n \times \gamma$, π is a 10D Gaussian and $\gamma = 0.7$. For reference, $\mathrm{KL}(\mu_5||\pi) = 10.2$. Right: Ratio of KL from *one* step of W-FR scheme to KL from continuous time WFR as a function of step size γ where π is a 2 component univariate Gaussian mixture.

Questions I

- When are the covariances negative (we have some conditions – but no results known in full generality).
- Continuous time/space very well studied. What is the best way to obtain discretisations?
- Interplay between splitting and numerical approximation.

Questions I

- When are the covariances negative (we have some conditions – but no results known in full generality).
- Continuous time/space very well studied. What is the best way to obtain discretisations?
- Interplay between splitting and numerical approximation.

Thank you!

Questions II

- Are other metrics possible (kernelised metrics? metrics outside W or FR?)
- Connections with other widely used sampling methods?

Questions II

- Are other metrics possible (kernelised metrics? metrics outside W or FR?)
- Connections with other widely used sampling methods?

Thank you!

References

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023.

Nicolas Chopin, Francesca Crucinio, and Anna Korba. A connection between tempering and entropic mirror descent. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8782–8800. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/chopin24a.html`.

Carles Domingo-Enrich and Aram-Alexandre Pooladian. An Explicit Expansion of the Kullback-Leibler Divergence along its Fisher-Rao Gradient Flow. *arXiv preprint arXiv:2302.12229*, 2023.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.