A connection between Sampling and Optimisation

Francesca R. Crucinio (King's College London)

Outline

1 Introduction

2 Variational Inference

3 Gradient Descent

4 Mirror Descent

Sampling

- **Aim 1:** sample from a probability distribution π on $\mathbb{R}^d$ and approximate expectations w.r.t. $\pi(x) = \eta(x)/\mathcal{Z}$ whose normalising constant might be unknown

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

- **Motivation:** compute posterior expectations in Bayesian inference
- **Aim 2:** estimate the unknown normalising constant $\mathcal{Z}$
- **Motivation:** model selection/parameter inference

Sampling as optimisation over distributions

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

where $\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log(\mu/\pi)\mathrm{d}\mu$ denotes the Kullback–Leibler divergence.

- Variational Inference (Blei et al., 2017)
- Algorithms based on the Langevin diffusion (Jordan et al., 1998)
- Stein Variational Gradient Descent (SVGD; Liu (2017))
- Algorithms based on tempering (Chopin et al. (2024) and Domingo-Enrich and Pooladian (2023))

Outline

1 Introduction

2 Variational Inference

3 Gradient Descent

4 Mirror Descent

Variational Inference[1]

In variational inference (or variational Bayes) we solve

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

Usually Ω corresponds to a certain **parametric family** (e.g. multivariate Gaussian distributions).

Optimisation happens at the parameter level, hence in $\mathbb{R}^d$.

[1] D. Blei et al, *Variational inference: A review for statisticians, JASA 2017*

Outline

1 Introduction

2 Variational Inference

3 Gradient Descent

- Langevin based algorithms
- Stein's algorithms

4 Mirror Descent

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

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

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

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

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

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla_{\mathcal{M}} \mathcal{F}(\mu_t)\right)$$

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

In the case of $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ we obtain

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla_{\mathcal{M}} \mathrm{KL}(\mu_t|\pi)\right).$$

1 Introduction

2 Variational Inference

3 Gradient Descent

- Langevin based algorithms
- Stein's algorithms

4 Mirror Descent

- Tempering

Wasserstein distance

Restrict to

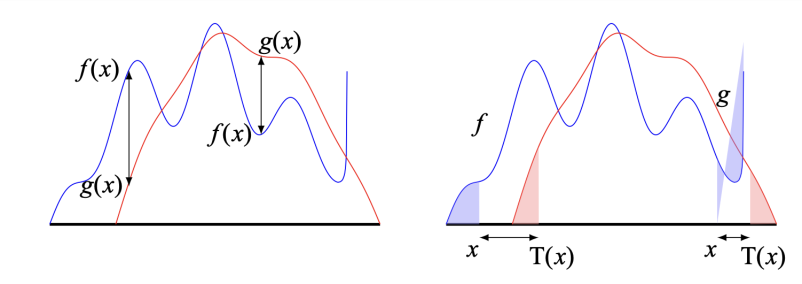
$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|^2 d\mu(x) < +\infty\}$$

and define the W_2 distance as

$$W_2(\mu, \nu) = \left(\inf_{\gamma \in \mathsf{T}(\mu,\nu)} \int \|x - y\|^2 d\gamma(x, y) \right)^{1/2}$$

where $\mathsf{T}(\mu, \nu)$ denotes the set of joint distributions which have μ and ν as marginals.

W_2 vs L^2



2

[2] F. Santambrogio, *Euclidean, Metric, and Wasserstein Gradient Flows: an overview*, Bulletin of Mathematical Sciences, 2017

Gradient descent w.r.t. W_2[3]

We have $\nabla_{W_2} \mathrm{KL}(\mu_t|\pi) = \nabla \log\left(\frac{\mu_t}{\pi}\right)$ from which we obtain the **Wasserstein gradient flow PDE**

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

[3] R, Jordan et al, *The variational formulation of the Fokker–Plank equation*, SIAM Mathematical Analysis 1998

Gradient descent w.r.t. W_2[3]

We have $\nabla_{W_2} \mathrm{KL}(\mu_t|\pi) = \nabla \log\left(\frac{\mu_t}{\pi}\right)$ from which we obtain the **Wasserstein gradient flow PDE**

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

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**.

[3] R, Jordan et al, *The variational formulation of the Fokker–Plank equation*, SIAM Mathematical Analysis 1998

Simple Euler–Maruyama discretisation leads to the **Unadjusted Langevin Algorithm** (ULA; Durmus and Moulines (2019))

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

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; Durmus and Moulines (2019))

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

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; Roberts and Tweedie (1996))
- Random walk Metropolis (RWM; Roberts et al. (1997))

1 Introduction

2 Variational Inference

3 Gradient Descent

- Langevin based algorithms
- Stein's algorithms

4 Mirror Descent

- Tempering

Stein geometry

A discrepancy measure based on Stein's identity

$$D_S(\mu, \nu) = \max_{\|\phi\|_{\mathcal{H}} \leq 1} \left\{ \mathbb{E}_{X\sim\mu}[\nabla \log f_\nu(X)^T \phi(X) + \nabla \cdot \phi(X)] \right\},$$

$\mathcal{H}$ is a reproducible kernel Hilbert space associated with a kernel k (e.g. gaussian).

Gradient descent w.r.t. Stein discrepancy

We have

$$\nabla_{\text{Stein}} \mathrm{KL}(\mu_t|\pi) = \int k(x,\cdot)\nabla \log\left(\frac{\mu_t}{\pi}(x)\right) d\mu_t(x).$$

The corresponding nonlinear PDE is

$$\begin{aligned} \partial_t \mu_t(x) &= \mathrm{div}\left(\mu_t(x) \int k(x,\cdot)[\nabla \mu_t + \mu_t \nabla \log \pi]\right) \\ &= -\mathrm{div}\left(\mu_t(x) \int [\nabla \log \pi(x) k(x,\cdot)] + \nabla_1 k(x,\cdot)] d\mu_t(x)\right) \end{aligned}$$

using integration by parts.

Stein variational gradient descent (SVGD)

We can approximate the behaviour of the nonlinear PDE with an **interacting particle system**

$$dX_t^i = 1/N \sum_{j=1}^{N} \left[k(X_t^i, X_t^j) \nabla \log \pi(X_t^j) - \nabla_1 k(X_t^j, X_t^i) \right]$$

for $i = 1, \ldots, N$.

An Euler–Maruyama discretisation gives the algorithm.

Outline

1 Introduction

2 Variational Inference

3 Gradient Descent

4 Mirror Descent

- Tempering

Mirror descent in Euclidean space

Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^+$ be a functional on $\mathbb{R}^d$. **Mirror Descent** proceeds iteratively solving

$$z_{n+1} = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{F}(z_n) + \langle \nabla \mathcal{F}(z_n), z - z_n \rangle + (\gamma_{n+1})^{-1} B_\phi(z|z_n) \right\} .$$

- $(\gamma_n)_{n \geq 0}$ is a sequence of step-sizes
- $B_\phi(z_1|z_2) = \phi(z_1) - \phi(z_2) - \langle \nabla \phi(z_2), z_1 - z_2 \rangle$ for some positive and convex ϕ is the **Bregman divergence**

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

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. **Mirror Descent** proceeds iteratively solving (Aubin-Frankowski et al., 2022)

$$\mu_{n+1} = \underset{\mu \in \mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}} \left\{ \mathcal{F}(\mu_n) + \langle \nabla \mathcal{F}(\mu_n), \mu - \mu_n \rangle + (\gamma_{n+1})^{-1} B_\phi(\mu|\mu_n) \right\}. \quad (1)$$

- $(\gamma_n)_{n \geq 0}$ is a sequence of step-sizes
- $B_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - \langle \nabla \phi(\mu), \nu - \mu \rangle$ for some positive and convex ϕ is the **Bregman divergence**
- $\langle \nabla \mathcal{F}(\nu), \xi \rangle = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\nu + \epsilon \xi) - \mathcal{F}(\nu))$ is the **first variation** of $\mathcal{F}$

Entropic mirror descent (MD)

Using the first order conditions of (1) we obtain the dual iteration

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\gamma_{n+1}\nabla\mathcal{F}(\mu_n).$$

In the case $B_\phi(\nu|\mu) = \mathrm{KL}(\nu|\mu)$, $\nabla\phi(\mu) = \log\mu$ and we have the following multiplicative update named **entropic mirror descent**:

$$\mu_{n+1} \propto \mu_n e^{-\gamma_{n+1}\nabla\mathcal{F}(\mu_n)}.$$

Entropic mirror descent (MD)

Using the first order conditions of (1) we obtain the dual iteration

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\gamma_{n+1}\nabla\mathcal{F}(\mu_n).$$

In the case $B_\phi(\nu|\mu) = \mathrm{KL}(\nu|\mu)$, $\nabla\phi(\mu) = \log\mu$ and we have the following multiplicative update named **entropic mirror descent**:

$$\mu_{n+1} \propto \mu_n e^{-\gamma_{n+1}\nabla\mathcal{F}(\mu_n)}.$$

If $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$, $\nabla\mathcal{F}(\mu) = \log(\frac{\mu}{\pi})$ and we obtain entropic mirror descent on the KL:

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})}\pi^{\gamma_{n+1}}.$$

1 Introduction

2 Variational Inference

3 Gradient Descent

- Langevin based algorithms
- Stein's algorithms

4 Mirror Descent

- Tempering

Tempering/Annealing

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

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}},$$

where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = 1$.

- Parallel Tempering (Geyer, 1991)
- Annealed Importance Sampling (Neal, 2001)
- Sequential Monte Carlo samplers (Del Moral et al., 2006)
- Termodynamic Integration (Gelman and Meng, 1998)

Connection between Tempering and MD

MD

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})} \pi^{\gamma_{n+1}}$$

Tempering

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$$

are equivalent if

$$\lambda_n = 1 - \prod_{k=1}^{n} (1 - \gamma_k).$$

Connection between Tempering and MD

MD

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})} \pi^{\gamma_{n+1}}$$

Tempering

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$$

are equivalent if

$$\lambda_n = 1 - \prod_{k=1}^{n} (1 - \gamma_k).$$

The connection between MD and tempering allows us to obtain explicit **rates of convergence** for the tempering iterates:

$$\mathrm{KL}(\mu_n | \pi) \leq \frac{\prod_{k=1}^{n}(1-\gamma_k)}{\gamma_1} \mathrm{KL}(\pi | \mu_0) = \frac{1-\lambda_n}{\lambda_1} \mathrm{KL}(\pi | \mu_0).$$

Choice of tempering sequence

The tempering iterates $\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}$ can be written in exponential family form

$$\mu_{n+1}(x) \equiv \mu_{\lambda_{n+1}}(x) \propto \mu_0 \exp\left\{\lambda_{n+1} s(x)\right\}$$

where $s(x) := \log \pi(x)/\mu_0(x)$.

We can compute the f-divergence between two successive iterates

$$\int \mu_\lambda f(\mu_{\lambda'}/\mu_\lambda) = \frac{f''(1)\mathrm{I}(\lambda)}{2} \times (\lambda' - \lambda)^2 + \mathcal{O}\left((\lambda' - \lambda)^3\right),$$

where $\mathrm{I}(\lambda) = \mathrm{Var}_{\mu_\lambda}\left[s(X)\right]$ is the Fisher information.

Adaptive choice of tempering

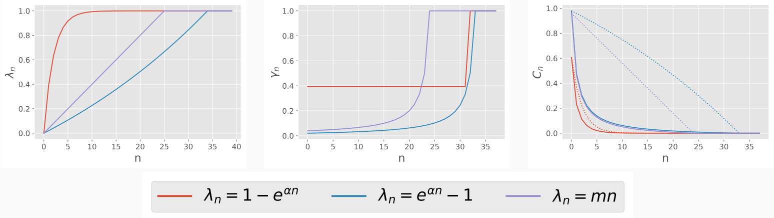
Intuitively, the distance between successive iterates should be small and constant. This suggests the following recipe to choose successive λ_n values:

$$\lambda_n - \lambda_{n-1} = c\mathrm{I}(\lambda_{n-1})^{-1/2} \tag{2}$$

Adaptive choice of tempering

Intuitively, the distance between successive iterates should be small and constant. This suggests the following recipe to choose successive λ_n values:

$$\lambda_n - \lambda_{n-1} = c\mathrm{I}(\lambda_{n-1})^{-1/2} \tag{2}$$



Algorithms

$$\mu_{n+1} \propto q_n \exp(-\gamma_n g_n)$$

where g_n is an approximation of the gradient of the KL objective $\log(\mu_n/\pi)$; and q_n is an approximation of μ_n.

We focus on algorithms which use:

- importance weights corresponding to $\exp(-\gamma_n g_n)$
- mixtures corresponding to q_n

Sequential Monte Carlo (SMC) samplers

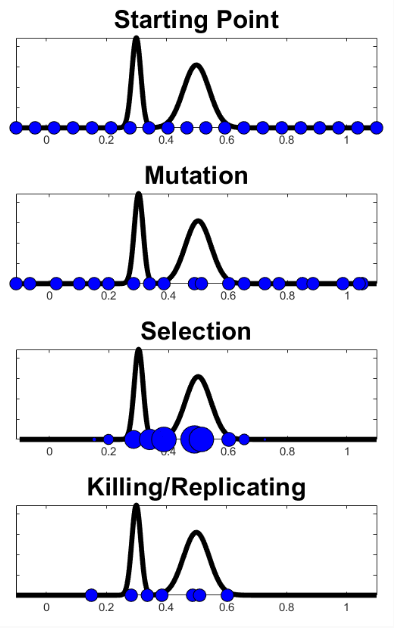
In SMC (Del Moral et al., 2006), the mirror descent iterate at time n is approximated by $q_n^{\mathrm{SMC}}(x) = \sum_{i=1}^{N} W_n^i \delta_{X_n^i}(x)$

- $\{X_n^i, W_n^i\}_{i=1}^{N}$ weighted particle set with

$$W_n(x) = \left(\frac{\pi(x)}{\mu_0(x)}\right)^{\lambda_n - \lambda_{n-1}} = \left(\frac{\pi(x)}{\mu_{n-1}(x)}\right)^{\gamma_n}. \tag{3}$$

- at each iteration a new N-particle set is resampled using W_n^i and a μ_n-invariant Markov kernel

Sequential Monte Carlo (SMC) samplers: basic idea



Adaptive strategies

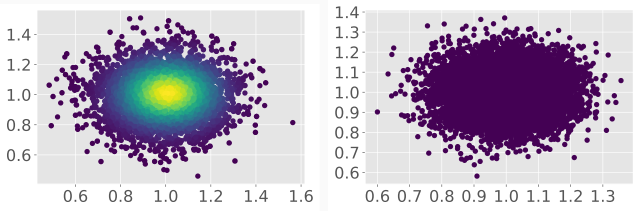
The SMC literature offers an easy way to tune the stepsize/tempering sequence adaptively: aim for iterates which keep constant

$$\mathrm{ESS}_n(\lambda) := 1/\sum_{i=1}^{N}(W_n^i)^2.$$

1. easy and inexpensive to approximate with particle cloud
2. approximates the χ^2 divergence $\chi^2(\mu_{\lambda'}|\mu_\lambda) \approx \frac{N}{\mathrm{ESS}_n(\lambda)} - 1$

Example

Approximations of $\pi = \mathcal{N}(1_d, 0.1^2 Id)$ from $\mu_0 = \mathcal{N}(0_d, Id)$.



Left: Adaptive SMC, Right: Fixed γ SMC.

Conclusions

- the connection between mirror descent (MD) and tempering justifies tempering from an optimisation point of view and provides the MD literature with several classes of algorithms (which are very well-studied!)
- opens the door to extensions of tempering through the use of other divergences
- gives a strategy to select γ/λ adaptively

Conclusions

- the connection between mirror descent (MD) and tempering justifies tempering from an optimisation point of view and provides the MD literature with several classes of algorithms (which are very well-studied!)
- opens the door to extensions of tempering through the use of other divergences
- gives a strategy to select γ/λ adaptively

Thank you!

Bibliography i

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Nicolas Chopin, Francesca Crucinio, and Anna Korba. A connection between tempering and entropic mirror descent. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=BtbijvkWLC`.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

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.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A): 2854–2882, 2019.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998.

Charles J Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramides, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

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.

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf`.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth–death dynamics for sampling: global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731, 2023.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.

Gradient flow with the Fisher-Rao geometry

Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ be a functional on $\mathcal{P}(\mathbb{R}^d)$. The gradient flow of F w.r.t. the Fisher-Rao geometry

$$d_H(\nu_1, \nu_2)^2 = 4 \int (\sqrt{\nu_1} - \sqrt{\nu_2})^2$$

can be written as (Domingo-Enrich and Pooladian, 2023; Lu et al., 2023)

$$\frac{\partial \mu_t}{\partial t} = -\mu_t \nabla \mathcal{F}(\mu_t), \text{ hence, } \frac{\partial \log(\mu_t)}{\partial t} = -\nabla \mathcal{F}(\mu_t).$$

Mirror descent (and tempering!) can be obtained as an Euler discretisation of the FR gradient flow.