Properties and limitations of geometric tempering for gradient flow dynamics

Anonymous authors
Paper under double-blind review

Abstract

We consider the problem of sampling from a probability distribution π. It is well known that this can be written as an optimisation problem over the space of probability distributions in which we aim to minimise the Kullback–Leibler divergence from π. We consider the effect of replacing π with a sequence of moving targets $(\pi_t)_{t\geq 0}$ defined via geometric tempering on the Wasserstein and Fisher–Rao gradient flows. We show that replacing the target distribution with a geometric mixture of initial and target distribution does not lead to a convergence speed up.

1 Introduction

Sampling from a target probability distribution whose density is known up to a normalisation constant is a fundamental task in computational statistics and machine learning. A natural way to formulate this task is optimisation of a functional measuring the dissimilarity to the target probability distribution, typically the Kullback–Leibler (KL) divergence due to the fact that the KL does not require knowledge of the normalisation constant (Chen et al. (2023, Theorem 4.1) and Crucinio (2025b)).

When the space over which the optimisation is carried out is the Wasserstein space, i.e. probability distributions with bounded second moments equipped with the Wasserstein-2 distance (Ambrosio et al., 2008), the resulting gradient descent scheme is referred to as Wasserstein (W) gradient flow.

A second class of gradient flows which has gained popularity in recent years is that of Fisher–Rao (FR) gradient flows, i.e. gradient descent w.r.t. the Fisher–Rao or Hellinger–Kakutani distance. Contrary to the Wasserstein metric, which leads to a diffusive behaviour, the Fisher–Rao metric leads to birth-death dynamics in which mass is created or removed (Lu et al., 2019; 2023).

Combinining the W and the FR gradient flows by making use of the Wasserstein–Fisher–Rao metric (Gallouët & Monsaingeon, 2017) leads to the Wasserstein–Fisher–Rao (WFR) gradient flow. The WFR flow combines the diffusive behaviour of the W gradient flow with the FR flow and enjoys more favourable convergence properties, combining the rate of convergence of the W flow with that of the FR flow (Domingo-Enrich & Pooladian, 2023; Lu et al., 2019); its numerical approximation can considerably speed up the convergence of gradient based methods for multimodal targets (Lu et al., 2019; 2023).

In this work we consider *tempered* versions of the Wasserstein, Fisher–Rao and Wasserstein–Fisher–Rao PDEs obtained by considering a time varying target $\pi_t \propto \pi^{\lambda_t}\mu_0^{1-\lambda_t}$ for $\lambda_t : \mathbb{R}_+ \to [0,1]$. We show that in the continuous time setting, tempering the target π does not improve convergence speed for the Wasserstein, the Fisher–Rao and the Wasserstein–Fisher–Rao flow (Proposition 1 and 2). Similar results hold for the discrete time dynamics (Proposition 3 and equation 14). We explore the gradient flow structure of these tempered dynamics and obtain convergence guarantees for the joint optimisation w.r.t. μ and λ using recent advances on optimisation on the product space $\mathcal{P}(\mathbb{R}^d) \times [0,1]$ (Caprio et al., 2025+; Crucinio, 2025a). This point of view allows to not fix λ_t in advance as is normally done with tempering dynamics but to learn it as time progresses.

Our theoretical results complement those shown in Chehab et al. (2025) for the W flow and confirm that tempering does not necessarily improve the dynamics of Wasserstein, Fisher–Rao and Wasserstein–Fisher–Rao flows.

Notation We define some notation that will be used throughout the manuscript. We endow $\mathbb{R}^d$ with the Borel σ-field $\mathcal{B}(\mathbb{R}^d)$ with respect to the Euclidean norm $\|\cdot\|$ where d is clear from context. We denote by $C^1(\mathbb{R}^d)$ the set of continuously differentiable functions, for all $f \in C^1(\mathbb{R}^d)$ we denote the gradient by ∇f. Throughout this manuscript we denote the target by π and assume $\pi = \exp(-V_\pi)$ for some $V_\pi \in C^1(\mathbb{R}^d)$. The initial distribution for the tempering iterates is denoted by $\mu_0 = \exp(-V_0)$ with $V_0 \in C^1(\mathbb{R}^d)$. Given π, μ_0 we define $\pi_t \propto \pi^{\lambda_t}\mu_0^{1-\lambda_t}$ for $\lambda_t : \mathbb{R}_+ \to [0,1]$.

2 Tempered PDEs for sampling

We consider a class of PDEs in which the target π is replaced by a sequence of moving targets $(\pi_t)_{t\geq 0}$. We refer to these PDEs as *tempered* PDEs. The original non-tempered dynamics can be obtained by replacing π_t with π throughout. Samplers based on tempered dynamics have become popular in the literature with the hope that the introduction of moving targets will improve the convergence of the sampler (Nüsken et al., 2024; Albergo & Vanden-Eijnden, 2024; Chehab et al., 2025; Guo et al., 2025).

We focus here on the sequence $(\pi_t)_{t\geq 0}$ defined by the geometric tempering iteration $\pi_t \propto \pi^{\lambda_t}\mu_0^{1-\lambda_t}$ where the tempering schedule $\lambda_t : \mathbb{R}_+ \to [0,1]$ is non-decreasing in t and weakly differentiable. This has been popularised in the sampling literature through its use in sequential Monte Carlo (SMC; Del Moral et al. (2006)), annealed importance sampling (AIS; Neal (2001)), parallel tempering (PT; Geyer (1991)) and the homotopy method (Daum & Huang, 2008).

The tempering sequence λ_t is fixed a priori. Popular options include the linear rate $\lambda_s = s/t$, the exponential rate $\lambda_s = 1 - e^{-\alpha s}$ for $\alpha > 0$ and the optimal rate for log-concave targets in Chehab et al. (2025, Section 4.2). We empirically compare this sequences in Section 4.

2.1 Tempered Wasserstein dynamics

We start by considering the tempered Wasserstein flow of Chehab et al. (2025). The corresponding PDE is given by

$$\partial_t \mu_t(x) = \nabla \cdot \left(\mu_t(x) \nabla \log \left(\frac{\mu_t(x)}{\pi_t(x)} \right) \right) \quad t \in [0, \infty). \tag{1}$$

This PDE is not a gradient flow of $D_{\mathrm{KL}}(\mu||\pi)$ and thus decay of $D_{\mathrm{KL}}(\mu||\pi)$ along the flow needs to be checked directly (e.g. Chehab et al. (2025, Theorem 1)).

We now analyse the decay of the Kullback–Leibler divergence along the tempered W PDE and compare it with its standard non-tempered counterpart. We obtain a result similar to Chehab et al. (2025, Theorem 1) (see Appendix A for a proof). Our result carries the same message as Chehab et al. (2025, Theorem 1) but allows for an easier comparison between tempered and standard PDEs by splitting the contribution to the decay of KL into the classical rate for standard W flows and a bias term.

In order to obtain our results we make the following assumptions on π, μ_0. These assumptions are strong but somewhat classical in the study of convergence of Langevin based samplers (e.g., Durmus et al. (2019); Chehab et al. (2025)).

Assumption 1. *The negative log-densities V_π, V_0 have Lipschitz continuous gradients with Lipschitz constants L_π, L_0, respectively. In addition, there exists $c_\pi, c_0 > 0$ such that*

$$\langle x - x', \nabla V_\pi(x) - \nabla V_\pi(x') \rangle \geq c_\pi \|x - x'\|^2, \qquad \langle x - x', \nabla V_0(x) - \nabla V_0(x') \rangle \geq c_0 \|x - x'\|^2.$$

Assumption 1 implies that π satisfies a log-Sobolev inequality (e.g., Chewi et al. (2024)):

$$D_{\mathrm{KL}}(\mu||\pi) \leq \frac{c_\pi^{-1}}{2} \int \mu \|\nabla \log \frac{\mu}{\pi}\|^2. \tag{2}$$

In addition, Assumption 1 implies the dissipativity condition

$$\langle x, \nabla V_\pi(x)\rangle \geq a_\pi \|x\|^2 - b_\pi,$$

for some $a_\pi > 0$ and $b_\pi > 0$. Assumption 1 also implies that $\nabla V_t := (1-\lambda_t)\nabla V_0 + \lambda_t \nabla V_\pi$ is Lipschitz continuous with Lipschitz constant $L_t = (1-\lambda_t)L_0 + \lambda_t L_\pi$ and strongly convex with constant $c_t \geq (1-\lambda_t)c_0 + \lambda_t c_\pi$.

Proposition 1 (Tempered W)**.** *Under Assumption 1, the tempered W PDE equation 1 reduces the Kullback–Leibler divergence at rate*

$$D_{\mathrm{KL}}(\mu_t||\pi) \leq e^{-2tc_\pi} D_{\mathrm{KL}}(\mu_0||\pi) + A\int_0^t e^{-2(t-s)c_\pi}(1-\lambda_s)ds, \tag{3}$$

where $A < \infty$ is a constant which only depends on the dissipativity and Lipschitz properties of V_π, V_0. The convergence rate of the standard W PDE is recovered for $\lambda_s \equiv 1$.

The first term corresponds to the rate of the classical W PDE, to which a bias term is added due to the fact that π_t approaches π only as $t \to \infty$. The bias term only disappears for $\lambda_s \equiv 1$, corresponding to the standard Wasserstein gradient flow.

2.2 Tempered Fisher–Rao dynamics

Given the tempering sequence π_t, one can also define the tempered FR PDE

$$\partial_t \mu_t(x) = \mu_t(x)\left(\log\left(\frac{\pi_t(x)}{\mu_t(x)}\right) - \mathbb{E}_{\mu_t}\left[\log\left(\frac{\pi_t}{\mu_t}\right)\right]\right). \tag{4}$$

Similarly to equation 1, the PDE equation 4 is not a gradient flow of $D_{\mathrm{KL}}(\mu||\pi)$.

Using the fact that

$$\frac{\partial e^t \log \mu_t(x)}{\partial t} = e^t \log \pi_t(x) - e^t \mathbb{E}_{\mu_t}\left[\log\left(\frac{\pi_t}{\mu_t}\right)\right]$$

and the expression for π_t we can obtain an analytic expression for the solution of the tempered FR PDE

$$\log \mu_t(x) = \log \mu_0(x)(\int e^{s-t}(1-\lambda_s)ds + e^{-t}) + \log \pi(x)\int e^{s-t}\lambda_s ds - \int_0^t e^{s-t}\mathbb{E}_{\mu_s}\left[\log\left(\frac{\pi_s}{\mu_s}\right)\right]ds.$$

It follows that

$$\mu_t \propto \mu_0^{e^{-t}+\int_0^t e^{s-t}(1-\lambda_s)ds}\pi^{\int_0^t e^{s-t}\lambda_s ds}. \tag{5}$$

We now show that equation 4 reduces $D_{\mathrm{KL}}(\mu||\pi)$. We consider the same assumptions required to obtain the convergence results for the standard FR flow in Chen et al. (2023, Theorem 4.1).

Assumption 2. *Assume that μ_0, π have bounded second moments $\int \|x\|^2\mu_0(x)dx \leq B$, $\int \|x\|^2\pi(x)dx \leq B$ and*

$$e^{-M(1+\|x\|^2)} \leq \frac{\mu_0(x)}{\pi(x)} \leq e^{M(1+\|x\|^2)} \tag{6}$$

for some $M > 0$.

Under this assumption, we can show that equation 4 reduces $D_{\mathrm{KL}}(\mu||\pi)$ at a rate which is always worse than that of the standard FR PDE (see Appendix A for a proof).

Proposition 2 (Tempered FR)**.** *Assume 2. The tempered FR PDE equation 4 reduces the Kullback–Leibler divergence at rate*

$$D_{\mathrm{KL}}(\mu_t||\pi) \leq M(1-\int_0^t e^{s-t}\lambda_s ds)\left[2+B+B\exp\left(M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds)\right)\right]. \tag{7}$$

The convergence result for the standard FR flow in Chen et al. (2023, Theorem 4.1) is recovered setting $\lambda_s \equiv 1$.

Since $1-\int_0^t e^{s-t}\lambda_s ds - e^{-t} = \int_0^t e^{s-t}(1-\lambda_s)ds \geq 0$, we conclude that the tempered FR PDE does not improve on the convergence rate of the standard FR PDE given by $Me^{-t}\left[2+B+B\exp\left(M(1+B)e^{-t}\right)\right]$. To further see this consider the evolution equation 5 and compare with the evolution along the standard FR PDE $\mu_t \propto \mu_0^{e^{-t}}\pi^{1-e^{-t}}$. We have $\mu_t \equiv \pi$ only for $t \to +\infty$, and since $1-\int_0^t e^{s-t}\lambda_s ds \geq e^{-t}$, it is clear that the tempered FR evolution moves more slowly than the standard untempered one.

2.3 Tempered Wasserstein–Fisher–Rao dynamics

Given the tempered W and tempered FR PDE one can naturally define a tempered WFR PDE

$$\partial_t \mu_t = \nabla\cdot\left(\mu_t(x)\nabla\log\left(\frac{\mu_t(x)}{\pi_t(x)}\right)\right) + \mu_t(x)\left(\log\left(\frac{\pi_t(x)}{\mu_t(x)}\right) - \mathbb{E}_{\mu_t}\left[\log\left(\frac{\pi_t}{\mu_t}\right)\right]\right). \tag{8}$$

By defining the derivative of $D_{\mathrm{KL}}(\mu||\pi)$ along the tempered W flow equation 1 and that along the tempered FR flow equation 4 as

$$\frac{d}{dt}D_{\mathrm{KL}}(\mu_t^{\text{T-W}}||\pi) := -\int \mu_t\|\nabla\log\frac{\mu_t}{\pi}\|^2 + (1-\lambda_t)\int\log\frac{\mu_t}{\pi}\nabla\cdot\left(\mu_t\nabla\log\left(\frac{\pi}{\mu_0}\right)\right)$$
$$\frac{d}{dt}D_{\mathrm{KL}}(\mu_t^{\text{T-FR}}||\pi) := -\mathrm{var}_{\mu_t}\left(\log\frac{\mu_t}{\pi}\right) - (1-\lambda_t)\mathrm{cov}_{\mu_t}\left(\log\frac{\mu_t}{\pi}, \log\frac{\pi}{\mu_0}\right)$$

we can write

$$\frac{d}{dt}D_{\mathrm{KL}}(\mu_t^{\text{T-WFR}}||\pi) = \frac{d}{dt}D_{\mathrm{KL}}(\mu_t^{\text{T-W}}||\pi) + \frac{d}{dt}D_{\mathrm{KL}}(\mu_t^{\text{T-FR}}||\pi),$$

showing that the W and FR geometries are indeed orthogonal as shown in (Gallouët & Monsaingeon, 2017).

Exploiting the fact that the time derivative of D_{KL} along the tempered WFR PDE factorises into two terms as in the case of the standard WFR flow, if μ_t follows the tempered WFR flow equation 8,

$$D_{\mathrm{KL}}(\mu_t|\pi) \leq \min\left\{D_{\mathrm{KL}}(\mu_t^{\text{T-W}}||\pi), D_{\mathrm{KL}}(\mu_t^{\text{T-FR}}||\pi)\right\}; \tag{9}$$

where an upper bound to $D_{\mathrm{KL}}(\mu_t^{\text{T-W}}||\pi)$ is given in equation 3 and one for $D_{\mathrm{KL}}(\mu_t^{\text{T-FR}}||\pi)$ in equation 7.

2.4 Multivariate Gaussian case

Our results as well as those in Chehab et al. (2025) suggest that the tempered dynamics do not improve on the gradient flow ones discussed in the previous section. However, our convergence results only obtain upper bounds for the decay of D_{KL} which are not necessarily tight. To confirm that the tempered PDEs do not improve over the standard ones, we consider the special case of transporting a Gaussian $\mu_0(x) = \mathcal{N}(x; m_0, \Sigma_0)$ to a target Gaussian $\pi(x) = \mathcal{N}(x; m_\pi, \Sigma_\pi)$, for which the W, FR, WFR, and their tempered versions can be solved analytically (see Appendix B).

In particular, observing that both the tempered W flow and the tempered FR flow preserve gaussianity, we have that $\mu_t(x) = \mathcal{N}(x; m_t, \Sigma_t)$ for each time $t \geq 0$ for equation 1, equation 4 and equation 8. The evolution of m_t and Σ_t is given by the following ODEs for the tempered W flow

$$\dot{m}_t = -(\lambda_t\Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})\left(m_t - ((1-\lambda_t)m_0\Sigma_\pi + \lambda_t m_\pi\Sigma_0)\left(\lambda_t\Sigma_0 + (1-\lambda_t)\Sigma_\pi\right)^{-1}\right)$$
$$\dot{\Sigma}_t = -(\lambda_t\Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})\Sigma_t - \Sigma_t(\lambda_t\Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1}) + 2I,$$

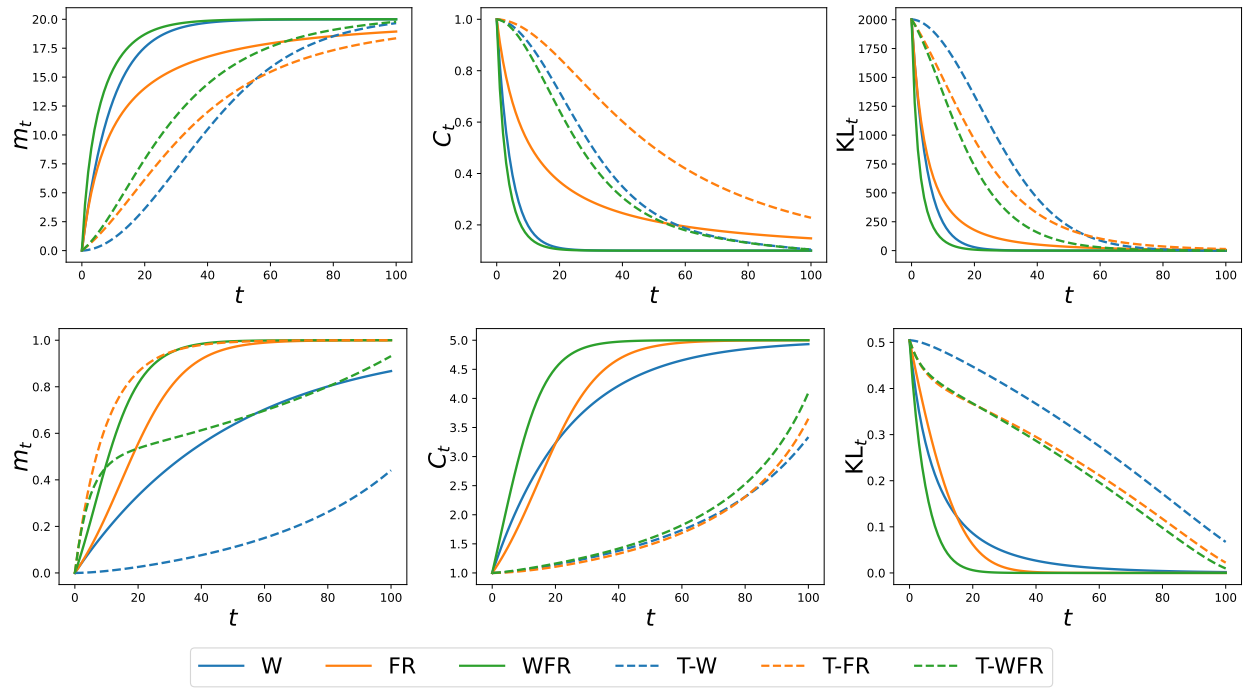

Figure 1: Evolution of mean, variance and D_{KL} along different PDE flows in the 1D Gaussian case with $\mu_0(x) = \mathcal{N}(x; 0, 1)$ and $\pi(x) = \mathcal{N}(x; 20, 0.1)$ (first row), $\pi(x) = \mathcal{N}(x; 1, 5)$ (second row).

and for the tempered FR flow by

$$\dot{m}_t = -\Sigma_t(\lambda_t \Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})(m_t - m_\pi)$$
$$\dot{\Sigma}_t = -\Sigma_t(\lambda_t \Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})\Sigma_t + \Sigma_t.$$

The standard (i.e. non-tempered) flows are obtained by setting $\lambda_t \equiv 1$ in the above equations. The tempered WFR flow is obtained by summing the r.h.s. of the W and the FR flows.

Figure 1 shows the evolution of mean, covariance and Kullback–Leibler divergence along different PDE flows in the case of 1D Gaussians. For the tempered PDEs we use the linear schedule $\lambda_s = s/t$. The initial distribution is a standard Gaussian and we consider two different targets: for target 1 (first row) we set $m_\pi = 20, \Sigma_\pi = 0.1$. In this case the W flow is faster than the FR one and the WFR flow behaves more like the W flow (Figure 1 top row). For the second target (second row) we have $m_\pi = 1, \Sigma_\pi = 5$; the FR flow is faster than the W flow for all sufficiently large t, and the WFR flow more closely follows the FR flow. Figure 1 confirms our theoretical results: the tempered PDEs result in a slower convergence to the target than the corresponding standard flows.

2.5 Discrete time convergence

Tempered unadjusted Langevin algorithms (Tempered ULA; Chehab et al. (2025)) provide an approximation of the tempered Wasserstein PDE equation 1 by using an Euler discretisation of the associated SDE

$$dX_t = \nabla \log \pi_t(X_t)dt + \sqrt{2}dB_t. \tag{10}$$

Denote by $\pi_n \propto \pi^{\lambda_n}\mu_0^{1-\lambda_n}$ obtained by considering a time discretisation $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq \lambda_{n+1} \leq \ldots$ of $\lambda_t : \mathbb{R}_+ \to [0, 1]$ with $\lambda_n \equiv \lambda_{t_n}$. The Tempered ULA dynamics for discretisation steps $(\gamma_n)_{n\geq 0}$ are given by

$$X_n = X_{n-1} + \gamma_n \nabla \log \pi_n(X_{n-1}) + \sqrt{2\gamma_n}\xi_n, \tag{11}$$

where $\xi_n \sim \mathcal{N}(0, \mathsf{Id})$. Denoting by μ_{n-1} the law of X_{n-1}, the update equation 11 corresponds to

$$\mu_n(x) = (4\pi\gamma_n)^{-d/2} \int \exp\left(-\frac{1}{4\gamma_n}\|x - y - \gamma_n \nabla \log \pi_n(y)\|^2\right) \mu_{n-1}(y)dy$$

where μ_n denotes the law of X_n.

Under Assumption 1 we can then obtain an equivalent result to Proposition 1 in discrete time (see Appendix C for a proof). This result is similar to Chehab et al. (2025, Theorem 3) and carries the same message. As in the case of Proposition 1 it allows for an easier comparison between tempered and standard PDEs by splitting the contribution to the decay of KL into the classical rate for standard W flows and a bias term.

Proposition 3. *Under Assumption 1, if* $\gamma_n \leq \min(1, \frac{a_\pi \wedge a_0}{2(L_\pi + L_0)^2}, \frac{c_n}{4L_n^2})$*, then*

$$D_{\mathrm{KL}}(\mu_n||\pi) \leq D_{\mathrm{KL}}(\mu_0||\pi)e^{-\sum_{k=1}^n \gamma_k c_k} + \sum_{k=0}^{n-1}(6\gamma_k^2 dL_k^2 + (1-\lambda_k)A')e^{-\sum_{i=1}^k \gamma_i c_i},$$

where A' *depends on the constants from Assumption 1, the second moment of the initial distribution* μ_0*, and linearly on the dimension.*

Similarly to Proposition 1, this result shows that the rate of convergence of tempered ULA is upper bounded by the rate of standard ULA (e.g. Vempala & Wibisono (2019, Theorem 1)) plus a bias term which disappears if $\lambda_k \equiv 1$.

Consider the tempered FR PDE equation 4; instead of using the exact evolution equation 5 consider an explicit time discretisation with discretisation steps $(\gamma_n)_{n\geq 0}$

$$\log(\mu_n) - \log(\mu_{n-1}) = -\gamma_n \left(\log\left(\frac{\mu_{n-1}}{\pi_{n-1}}\right) - \mathbb{E}_{\mu_{n-1}}\left[\log\left(\frac{\mu_{n-1}}{\pi_{n-1}}\right)\right]\right), \tag{12}$$

where $\pi_n \propto \pi^{\lambda_n}\mu_0^{1-\lambda_n}$ Observing that $\mathbb{E}_{\mu_{n-1}}\left[\log\left(\frac{\mu_{n-1}}{\pi_{n-1}}\right)\right]$ is just a normalising constant, we obtain the sequence of distributions

$$\mu_n \propto \mu_{n-1}^{(1-\gamma_n)}\pi_{n-1}^{\gamma_n} = \mu_{n-1}^{(1-\gamma_n)}\pi^{\lambda_{n-1}\gamma_n}\mu_0^{(1-\lambda_{n-1})\gamma_n}. \tag{13}$$

Proceeding recursively the above becomes

$$\mu_n \propto \mu_0^{1-\sum_{k=1}^n \gamma_k\lambda_{k-1}\prod_{j=k+1}^n(1-\gamma_j)}\pi^{\sum_{k=1}^n \gamma_k\lambda_{k-1}\prod_{j=k+1}^n(1-\gamma_j)}$$

which corresponds to one step of mirror descent with time step $\delta_n = \sum_{k=1}^n \gamma_k\lambda_{k-1}\prod_{j=k+1}^n(1-\gamma_j)$ applied to the Kullback–Leibler divergence (see Chopin et al. (2024)). The non-tempered FR dynamics correspond instead to mirror descent with $\delta_n = 1 - \prod_{k=1}^n(1-\gamma_k)$ wich correspond to the above when $\lambda_n \equiv 1$.

The fact that the mirror descent update reduces D_{KL} follows from Chopin et al. (2024, Proposition 1) with γ_n replaced by δ_n therein:

$$D_{\mathrm{KL}}(\mu_n||\pi) \leq (\delta_1)^{-1}\prod_{k=1}^n(1-\delta_k)D_{\mathrm{KL}}(\pi||\mu_0). \tag{14}$$

It follows that, since $\delta_n \leq 1$ for all $n \geq 1$, the tempered FR flow reduces KL at a geometric rate $\prod_{k=1}^n(1-\delta_k) \to 0$ as $n \to \infty$. To compare the tempered FR dynamics with the standard FR dynamics consider $1-\delta_n$, in the tempered case $1-\delta_n = \sum_{k=1}^n \gamma_k\lambda_{k-1}\prod_{j=k+1}^n(1-\gamma_j) \leq \sum_{k=1}^n \gamma_k\prod_{j=k+1}^n(1-\gamma_j) = \prod_{k=1}^n(1-\gamma_k)$. Thus tempering reduces the convergence rate in equation 14.

3 Gradient flow structure of tempered PDEs

The tempered PDEs described above do not correspond to gradient flows of $D_{\mathrm{KL}}(\mu||\pi)$. To identify a gradient flow structure, we consider the functional $\mathcal{F}(\mu,\lambda) : \mathcal{P}(\mathbb{R}^d) \times [0,1] \to \mathbb{R}$ defined as $\mathcal{F}(\mu,\lambda) := D_{\mathrm{KL}}(\mu||\pi_\lambda)$ where $\pi_\lambda \propto \mu_0^{1-\lambda}\pi^\lambda$ and optimisation now occurs over both μ and λ. In this case the sequence λ_t is not fixed a priori but learnt as time progresses.

3.1 Tempered Wasserstein dynamics

To define a gradient flow of $\mathcal{F}$ we need to consider a metric on $\mathcal{P}(\mathbb{R}^d) \times [0,1]$. To derive the tempered Wasserstein dynamics equation 1 from a gradient flow point of view, we consider the Wasserstein metric on $\mathcal{P}(\mathbb{R}^d)$ coupled with the Euclidean metric on $[0,1]$. The metric structure of this product space is investigated in Kuntz et al. (2023, Appendix A), which guarantees that $\mathcal{P}(\mathbb{R}^d) \times [0,1]$ is a metric space.

We follow Kuntz et al. (2023) and consider $\nabla\mathcal{F}(\mu,\lambda) = (\nabla_\mu \mathcal{F}(\mu,\lambda), \partial_\lambda \mathcal{F}(\mu,\lambda))$, where ∇_μ denotes the gradient w.r.t. μ in Wasserstein distance and ∂_λ the usual derivative in Euclidean space. Using Kuntz et al. (2023, Lemma 1 and Eq. (18)), we find

$$\nabla_\mu \mathcal{F}(\mu,\lambda) = \nabla \cdot \left(\mu \nabla \log \frac{\pi_\lambda}{\mu}\right), \qquad \partial_\lambda \mathcal{F}(\mu,\lambda) = -\int \mu(x) \partial_\lambda \log \pi_\lambda(x) dx.$$

The gradient flow dynamics are then given by

$$\begin{aligned} \partial_t \mu_t(x) &= -\nabla \cdot \left(\mu_t(x) \nabla \log \frac{\pi_{\lambda_t}(x)}{\mu_t(x)}\right) \\ \dot{\lambda}_t &= \int \mu_t(x)\, \partial_\lambda \log \pi_\lambda(x)|_{\lambda=\lambda_t}\, dx = \int \mu_t(x) \log \frac{\pi(x)}{\mu_0(x)} dx - \int \pi_{\lambda_t}(x) \log \frac{\pi(x)}{\mu_0(x)} dx \end{aligned} \tag{15}$$

as $\partial_\lambda \log \pi_\lambda(x) = \log \frac{\pi(x)}{\mu_0(x)} - \int \pi_\lambda(x) \log \frac{\pi(x)}{\mu_0(x)} dx$. It is now easy to see that the evolution of μ_t coincides with the tempered W PDE equation 1 provided that the appropriate λ_t is chosen.

Under Assumption 1, Caprio et al. (2025+, Theorem 1) guarantees that the Wasserstein gradient flow of $\mathcal{F}$ given by equation 15 converges exponentially fast.

Proposition 4. *Under Assumption 1, the Wasserstein gradient flow of $\mathcal{F}$ given by equation 15 satisfies*

$$0 \le \mathcal{F}(\mu_t, \lambda_t) \le \mathcal{F}(\nu, \ell) e^{-2t(c_0 \wedge c_\pi)} \tag{16}$$

where c_0, c_π denote the convexity constants of μ_0, π, respectively, and (ν, ℓ) is the initial condition of equation 15.

Proof. To apply Caprio et al. (2025+, Theorem 1) we check Assumption 1 and 2 therein as well as the extended log-Sobolev inequality (Eq. (11) therein).

The mapping $\lambda \mapsto \pi_\lambda = \mu_0^{1-\lambda} \pi^\lambda / Z_\lambda$ is twice continuously differentiable with derivatives

$$\partial_\lambda \log \pi_\lambda(x) = \log \frac{\pi(x)}{\mu_0(x)} - \int \pi_\lambda(x) \log \frac{\pi(x)}{\mu_0(x)} dx, \qquad \partial_\lambda^2 \log \pi_\lambda(x) = \mathrm{var}_{\pi_\lambda}\left(\frac{\pi}{\mu_0}\right)$$

which are finite quantities under Assumption 1. In addition, $\pi_\lambda(x) > 0$ for all λ, x if $\mu_0(x) > 0$ and $\pi(x) > 0$. The Lipschitz continuity in Assumption 1 then guarantees that Caprio et al. (2025+, Assumption 2) is satisfied.

To check the extended log-Sobolev inequality in Caprio et al. (2025+, Eq. (11)) we note that $\mathcal{F}^\star = 0$ and, since π_λ is strongly log-concave under Assumption 1,

$$\mathcal{F}(\mu,\lambda) = D_{\mathrm{KL}}(\mu || \pi_\lambda) \le \frac{1}{2c_\lambda} \int \|\nabla_x \log \frac{\mu(x)}{\pi_\lambda(x)}\|^2 \mu(x) dx \le \frac{1}{2(c_0 \wedge c_\pi)} \int \|\nabla_x \log \frac{\mu(x)}{\pi_\lambda(x)}\|^2 \mu(x) dx$$

where $c_\lambda \ge (1-\lambda)c_0 + \lambda c_\pi \ge c_0 \wedge c_\pi$ is the convexity constant of π_λ.

□

3.2 Tempered Fisher–Rao dynamics

We exploit the fact that mirror descent provides a time discretisation of the Fisher–Rao dynamics (Chopin et al., 2024) and consider directly the time-discretised solution of the tempered FR PDE equation 5.

Similarly to Crucinio (2025a), we consider a mirror descent scheme on the product space $\mathcal{P}(\mathbb{R}^d) \times [0,1]$. Given an initial $z_0 \in \mathcal{P}(\mathbb{R}^d) \times [0,1]$ and a sequence of step-sizes $(\gamma_n)_{n\geq 1}$, one can generate the sequence $(z_n)_{n\geq 0}$ as follows

$$z_{n+1} = \underset{z\in\mathcal{P}(\mathbb{R}^d)\times[0,1]}{\arg\min} \left\{\mathcal{F}(z_n) + \langle \delta\mathcal{F}(z_n), z - z_n\rangle + (\gamma_{n+1})^{-1} B_\phi(z|z_n)\right\}, \tag{17}$$

where B_ϕ denotes the Bregman divergence induced by a convex and continuously differentiable functional $\phi : \mathcal{P}(\mathbb{R}^d) \times [0,1] \to \mathbb{R}^+$, and $\delta\mathcal{F}(z)$ the first variation of $\mathcal{F}$ at $z = (\mu, \lambda)$ (see Appendix D for a derivation). Writing the first order conditions of equation 17, we obtain the dual iteration

$$\nabla\phi(z_n) - \nabla\phi(z_{n-1}) = -\gamma_n \delta\mathcal{F}(z_{n-1}).$$

Let B_h be any Bregman divergence over $[0,1]$ and denote $z = (\mu, \lambda)$. Then $B_\phi(z_1|z_2) = B_h(\lambda_1|\lambda_2) + D_{\mathrm{KL}}(\mu_1|\mu_2)$ is a Bregman divergence over $\mathcal{P}(\mathbb{R}^d) \times [0,1]$ with $\phi : (\mu, \lambda) \mapsto \int \log(\mu(x))\mu(x)dx + h(\lambda)$ as guaranteed by Crucinio (2025a, Proposition 1). Writing the first order conditions of equation 17 with Bregman divergence as above, we obtain the dual iteration

$$\begin{aligned} \nabla h(\lambda_n) - \nabla h(\lambda_{n-1}) &= \gamma_n \int \partial_\lambda \log \pi_\lambda(x)|_{\lambda=\lambda_{n-1}} \mu_{n-1}(x)dx \\ \log \mu_n(x) - \log \mu_{n-1}(x) &= -\gamma_{n-1}\left[\log \mu_{n-1}(x) - \log \pi_{\lambda_{n-1}}(x)\right]. \end{aligned} \tag{18}$$

The second component corresponds to the discretisation of the tempered FR PDE equation 12 up to the normalisation constant, and, since ∇h is bijective due to the convexity of h, we obtain the update for λ as follows

$$\lambda_n = (\nabla h)^{-1}\left(\nabla h(\lambda_{n-1}) + \gamma_n \int [\mu_n(x) - \pi_{n-1}(x)] \log \frac{\pi(x)}{\mu_0(x)} dx\right) \tag{19}$$

If B_h is the Euclidean distance, so that equation 19 becomes a gradient descent update, Assumption 1 guarantees that π_λ is both strongly log-concave and gradient Lipschitz continuous w.r.t. λ (see the argument in the proof of Proposition 4). We can then apply Crucinio (2025a, Proposition 1) to obtain the following result.

Proposition 5. *Under Assumption 1 if $h = \|\cdot\|^2/2$ and if $\gamma_n \leq \min(c_n, 1)^{-1}$ for all $n \geq 1$, we have*

$$0 \leq \mathcal{F}(\mu_n, \lambda_n) \leq (\gamma_1)^{-1} \prod_{k=1}^{n} (1 - \gamma_k \min(c_k, 1))[D_{\mathrm{KL}}(\pi||\nu) + \|\ell - 1\|^2],$$

where c_k is the convexity constant of π_k and (ν, ℓ) is the initial condition of equation 18.

4 Numerical Experiments

We now present some numerical experiments to support our results in the previous sections. Code to reproduce our experiments will be made available.

4.1 Algorithms

We briefly describe here the numerical approximations of the tempered W PDE and the tempered WFR PDE that we use in the following sections.

For the tempered W PDE we use the Tempered ULA dynamics equation 11. For the tempered WFR PDE we adapt the sequential Monte Carlo algorithm of Crucinio & Pathiraja (2025) to the use of tempered targets.

To obtain a time discretisation of the tempered WFR PDE we consider the solution operator of the tempered W flow and the tempered FR flow separately. Assume that for certain discrete time $n-1$ the solution of

the tempered WFR PDE is given by μ_{n-1}. Given μ_{n-1}, a discretisation of the tempered W flow for a time step γ_n leads to

$$\mu_{n-1/2}(x) \propto \int \exp\left(-\frac{1}{4\gamma_n}\|x - y - \gamma_n \nabla \log \pi_n(y)\|^2\right) \mu_{n-1}(y)dy,$$

which can be implemented using the tempered ULA equation 11. Using the exact solution of the tempered FR PDE equation 5 we have

$$\mu_n(x) \propto \mu_{n-1/2}(x)^{e^{-\gamma}+\int_0^\gamma e^{s-\gamma}(1-\lambda_s)ds} \pi(x)^{\int_0^\gamma e^{s-\gamma}\lambda_s ds} = \left(\frac{\pi(x)}{\mu_{n-1/2}(x)}\right)^{\int_0^\gamma e^{s-\gamma}\lambda_s ds} \mu_{n-1/2}(x), \tag{20}$$

where we used $e^{-t} + \int_0^t e^{s-t}(1-\lambda_s)ds = 1 - \int_0^t e^{s-t}\lambda_s ds$. Given $\mu_{n-1/2}$, we can obtain μ_n by importance sampling with weights $\left(\pi(x)/\mu_{n-1/2}(x)\right)^{\int_0^\gamma e^{s-\gamma}\lambda_s ds}$.

Alternating between one step of tempered W flow approximated via equation 11 and one step of tempered FR flow approximated via importance sampling we obtain an approximation of tempered WFR PDE through sampling-based procedures. A resampling step to avoid the weight degeneracy normally introduced by repeated importance sampling steps can also be added (Algorithm 1).

Algorithm 1 SMC-T-WFR.

Inputs: learning rates $(\gamma_n)_{n\geq 0}$, initial proposal μ_0.
Initialize: sample $\widetilde{X}_0^i \sim \mu_0$ and set $W_0^i = 1/N$ for $i = 1, \dots, N$.
for $n = 1, \dots, T$ **do**
  if $n > 1$ **then**
    Resample: draw $\{\widetilde{X}_{n-1}^i\}_{i=1}^N$ independently from $\{X_{n-1}^i, W_{n-1}^i\}_{i=1}^N$ and set $W_n^i = 1/N$ for $i = 1, \dots, N$.
  end if
  W flow: update $X_n^i = \bar{X}_{n-1}^i + \sqrt{2\gamma_n}\xi_{n-1/2}^i$, where $\bar{X}_{n-1}^i = \widetilde{X}_{n-1}^i + \gamma_n \nabla \log \pi(\widetilde{X}_{n-1}^i)$, $\xi_{n-1/2}^i \sim \mathcal{N}(0, \mathsf{Id})$ for $i = 1, \dots, N$.
  FR flow: compute and normalize the weights $W_n^i \propto w_n(X_n^i)$ for $i = 1, \dots, N$, where

$$w_n(x) = \left(\frac{\pi(x)}{N^{-1}\sum_{i=1}\mathcal{N}(x; X_{n-1}^i + \gamma_n \nabla \log \pi_n(X_{n-1}^i), 2\gamma_n \cdot \mathsf{Id})}\right)^{\int_0^{\gamma_n} e^{s-\gamma_n}\lambda_s ds}. \tag{21}$$

end for
Output: $\{X_n^i, W_n^i\}_{i=1}^N$.

Despite using the tempered unadjusted Langevin algorithm as proposal, Algorithm 1 does not coincide with the standard tempering SMC sampler (Chopin & Papaspiliopoulos, 2020, Chapter 17). To draw a connection between the two, consider the time discretisation of the tempered FR PDE in equation 13 with $\gamma_n \equiv 1$, $\mu_n \propto \pi^{\lambda_n}\mu_0^{(1-\lambda_n)}$. Using this sequence of distributions with the tempered unadjusted Langevin proposal equation 11 coincides with the commonly used tempering SMC and results in weights $w_n(x_n) = (\pi(x_n)/\mu_0(x_n))^{\lambda_n - \lambda_{n-1}}$ for sufficiently small γ (Dai et al., 2022).

4.2 Tempered and Standard PDEs

Proposition 1 shows that the tempered Wasserstein dynamics (T-W) do not improve on the standard Wasserstein gradient flows (W) in the case of strongly log-concave target π. A similar result holds for the WFR flow combining the rates in Proposition 1 with those in Proposition 2 as shown in equation 9. For bi-modal targets, which satisfy a log-Sobolev assumption but are not log-concave, (Chehab et al., 2025, Theorem 8) shows that the T-W flow has exponentially slower convergence than the standard W flow.

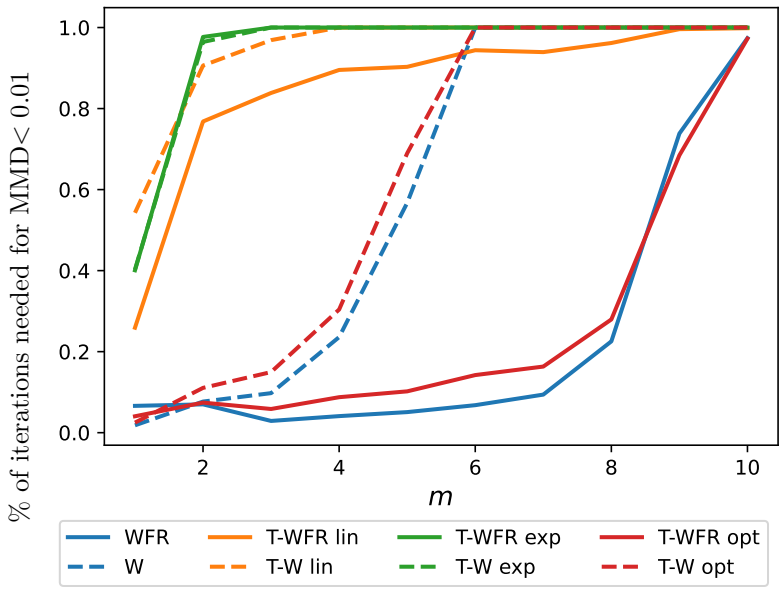

Figure 2: Convergence of approximations to π given by tempered and standard PDE flow with target the Gaussian mixture equation 22. Convergence is measured through the MMD with standard Gaussian kernel and is averaged over 50 replications. We plot the percent of iterations required to achieve MMD$<$ 0.01 for increasing m. We include standard W & WFR and tempered flows with $\lambda_s = s/t$, $\lambda_s = 1 - e^{-\alpha s}$, $\lambda_s = 1 - 1/(2+s)$.

We empirically show that the same is true for the tempered WFR flow. We consider the same setup of Chehab et al. (2025):

$$\mu_0(x) = \mathcal{N}(x; 0, 1), \qquad \pi(x) = \frac{1}{2}\mathcal{N}(x; 0, 1) + \frac{1}{2}\mathcal{N}(x; m, 1). \tag{22}$$

This target satisfies a log-Sobolev inequality (Schlichting, 2019) with $C_{\mathrm{LSI}} = (1 + (e^{m^2} + 1)/2)$.

For the tempered PDEs we consider the linear rate $\lambda_s = s/t$, the exponential rate $\lambda_s = 1 - e^{-\alpha s}$ for $\alpha = 0.01$ and the optimal rate of Chehab et al. (2025), $\lambda_s = 1 - 1/(2+s)$ (which follows from setting $\mu_0(x) = \mathcal{N}(x; 0, 1)$). We check convergence computing the MMD with standard Gaussian kernel.

Figure 2 shows the number of iteration needed to achieve MMD$<$ 0.01. For the W flows we run N parallel chains using the same number of iterations and discretisation steps as for the WFR flows. We see that the tempered dynamics do not perform better than the standard dynamics; the optimal rate performs similarly to the non-tempered dynamics for both W and WFR. WFR converges faster than the W flow for all choices of tempering regardless of the log-Sobolev constant C_{LSI}. The increased convergence speed of WFR comes at a higher computational cost; while ULA and its tempered variants have $\mathcal{O}(N)$ cost, WFR and Tempered WFR require the computation of the weights leading to an $\mathcal{O}(N^2)$ cost (Crucinio & Pathiraja, 2025).

4.3 Tempered SMC-WFR and Tempering SMC

We now compare two approximations of the tempered WFR flow: Algorithm 1 obtained using the analytic solution of the FR flow and the usual tempering SMC algorithm (e.g. Chopin & Papaspiliopoulos (2020)) which uses the sequence $\mu_n \propto \pi^{\lambda_n}\mu_0^{(1-\lambda_n)}$ and tempered ULA equation 11 as proposal kernel.

We empirically compare this two approximations on a 1D Gaussian example with $\mu_0(x) = \mathcal{N}(x; 0, 1)$ and $\pi(x) = \mathcal{N}(x; 20, 0.1)$ using the linear schedule $\lambda_s = s/t$. We take $\gamma \in [0.001, 0.1]$ and investigate how well the two approaches match the exact evolution of D_{KL} along the tempered WFR flow obtained using the moment

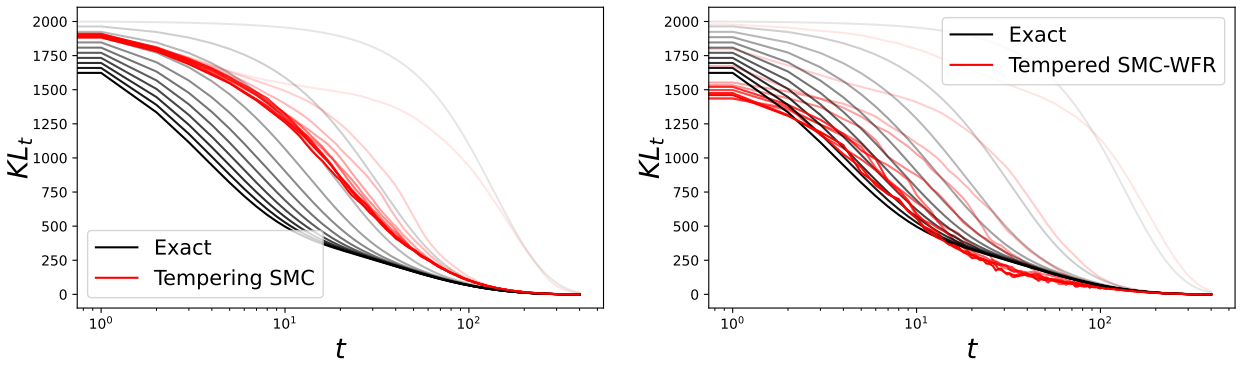

Figure 3: Evolution of D_{KL} along the tempered WFR PDE flow with linear schedule and two approximations via SMC. We have $\mu_0(x) = \mathcal{N}(x; 0, 1)$ and $\pi(x) = \mathcal{N}(x; 20, 0.1)$, γ ranges from 0.001 (lighter color) to 0.1 (darker color).

ODEs in Appendix B. Figure 3 shows that for small γ the two approximations are both close to the exact evolution along the tempered WFR PDE but as γ increases tempering SMC deviates from the continuous time evolution by slowing down convergence.

Despite the slower convergence, tempering SMC has some advantages over the approximation of the tempered WFR flow given by Algorithm 1: setting $\gamma_n \equiv 1$ in equation 13 results in weights which only involve π, μ_0 and thus can be computed in $\mathcal{O}(1)$ time against the $\mathcal{O}(N)$ time of equation 21. In addition, when using the sequence of distributions $\mu_n \propto \pi^{\lambda_n} \mu_0^{1-\lambda_n}$, one can set $\lambda_T = 1$ for some large T and therefore obtain π in finite time; this is not the case for the exact solution of the tempered FR PDE equation 5 for which convergence to π only occurs in the infinite time limit.

5 Discussion

We consider tempered dynamics and show that, both in continuous and in discrete time these dynamics do not improve the convergence speed of the standard, untempered ones. Even if our theoretical and empirical results (Section 4) suggest that tempered dynamics are not beneficial for sampling, we point out that the idea of tempering, i.e. to slowly move the initial distribution μ_0 to the target π via a geometric path, coincides with the unit time FR flow

$$\partial_t \mu_t(x) = \mu_t(x) \left(\log \left(\frac{\pi(x)}{\mu_0(x)} \right) - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\pi}{\mu_0} \right) \right] \right) \quad t \in [0, 1], \tag{23}$$

as highlighted in Chopin et al. (2024); Domingo-Enrich & Pooladian (2023).

Unit time FR flows are easily rescaled to the infinite time interval as shown in Domingo-Enrich & Pooladian (2023). We can therefore connect the tempering dynamics considered in this work with the Fisher–Rao flow in infinite time. It is intuitive that tempering the FR flow itself will not improve on its convergence (Proposition 2).

Combining the infinite time FR flow with the W flow leads to the WFR flow and is known to improve convergence (Lu et al., 2019). Therefore, changing the target in time using $\pi_t \propto \mu_0^{1-\lambda_t} \pi^{\lambda_t}$ only introduces a bias and does not speed up the WFR dynamics.

Acknowledgments

F.R.C. gratefully acknowledges the “de Castro" Statistics Initative at the *Collegio Carlo Alberto* and the *Fondazione Franca e Diego de Castro*. F.R.C. is supported by the Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA-INdAM).

S.P. gratefully acknowledges funding from UNSW Faculty of Science Research Grant and the Eva Mayr Stihl Foundation.

References

Michael S Albergo and Eric Vanden-Eijnden. Nets: A non-equilibrium transport sampler. *arXiv preprint arXiv:2410.02711*, 2024.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Rocco Caprio, Juan Kuntz, Samuel Power, and Adam M Johansen. Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities. *Journal of Machine Learning Research*, 2025+.

Omar Chehab, Anna Korba, Austin J Stromme, and Adrien Vacher. Provable convergence and limitations of geometric tempering for langevin dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=DZcmz9wU0i`.

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.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *Foundations of Computational Mathematics*, pp. 1–51, 2024.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.

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 (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8782–8800. PMLR, 21–27 Jul 2024.

Francesca R. Crucinio and Sahani Pathiraja. Sequential Monte Carlo approximations of Wasserstein–Fisher–Rao gradient flows. *arXiv preprint arXiv: 2506.05905*, 2025.

Francesca R Crucinio, Valentin De Bortoli, Arnaud Doucet, and Adam M. Johansen. Solving a class of Fredholm integral equations of the first kind via Wasserstein gradient flows. *Stochastic Processes and their Applications*, 173:104374, 2024. ISSN 0304-4149.

Francesca Romana Crucinio. A mirror descent approach to maximum likelihood estimation in latent variable models. *Journal of Computational and Graphical Statistics*, (just-accepted):1–19, 2025a.

Francesca Romana Crucinio. A note on the unique properties of the Kullback–Leibler divergence for sampling via gradient flows. *arXiv preprint arXiv:2507.04330*, 2025b.

Chenguang Dai, Jeremy Heng, Pierre E Jacob, and Nick Whiteley. An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association*, 117(539):1587–1600, 2022.

Fred Daum and Jim Huang. Particle flow for nonlinear filters with log-homotopy. In *Signal and Data Processing of Small Targets 2008*, volume 6969, pp. 414–425. SPIE, 2008.

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. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Thomas O Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

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

Wei Guo, Molei Tao, and Yongxin Chen. Provable Benefit of Annealed Langevin Monte Carlo for Non-log-concave Sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=P6IVIoGRRg`.

Juan Kuntz, Jen Ning Lim, and Adam M Johansen. Particle algorithms for maximum likelihood training of latent variable models. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin Sampling with Birth-death. *arXiv preprint arXiv: 1905.09863*, pp. 1–18, 2019.

Yulong Lu, Dejan Slepcev, and Lihan Wang. Birth-death dynamics for sampling: Global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731–5772, 2023. ISSN 0951-7715, 1361-6544.

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

Nikolas Nüsken, Francisco Vargas, Shreyas Padhy, and Denis Blessing. Transport Meets Variational Inference: Controlled Monte Carlo Diffusions. In *The Twelfth International Conference on Learning Representations: ICLR 2024*, 2024.

André Schlichting. Poincaré and log–Sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A Continuous time convergence

A.1 Tempered Wasserstein PDE

Consider the W PDE

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

where $\pi_t \propto \pi^{\lambda_t} \mu_0^{1-\lambda_t}$ and $\lambda_t : \mathbb{R}_+ \to [0,1]$, λ_t is non-decreasing in t and weakly differentiable.

A direct computation of the decay of $D_{\mathrm{KL}}(\mu_t || \pi)$ along this PDE gives

$$\begin{aligned} \frac{d}{dt} D_{\mathrm{KL}}(\mu_t || \pi) &= -\int \mu_t \|\nabla \log \frac{\mu_t}{\pi}\|^2 + (1-\lambda_t) \int \log \frac{\mu_t}{\pi} \nabla \cdot \left(\mu_t \nabla \log \left(\frac{\pi}{\mu_0} \right) \right) \\ &= -\int \mu_t \|\nabla \log \frac{\mu_t}{\pi}\|^2 + (1-\lambda_t) \int \mu_t \left[\Delta V_0 - \Delta V_\pi - \langle \nabla V_\pi, \nabla V_0 - \nabla V_\pi \rangle \right] \end{aligned}$$

where we used the expression for π_t and multiple applications of integrations by parts.

To proceed we recall that Assumption 1 implies the dissipativity condition

$$\langle x, \nabla V_\pi(x) \rangle \geq a_\pi \|x\|^2 - b_\pi, \tag{24}$$

with $a_\pi = c_\pi - 1/(2\epsilon) > 0$ and $b_\pi = \epsilon\|\nabla V_\pi(0)\|^2/2 > 0$ for $\epsilon > 1/(2c_\pi)$ (see, e.g., Crucinio et al. (2024, Appendix B.3) for a proof); and similarly for μ_0. A consequence of the dissipativity equation 24 is

$$\mathbb{E}_\pi[\|X\|^2] \leq \frac{1}{a_\pi}\mathbb{E}_\pi[\langle X, \nabla V_\pi(X)\rangle] + \frac{b_\pi}{a_\pi} = \frac{d + b_\pi}{a_\pi}. \tag{25}$$

In addition, V_π must be minimised at some point $z_0 \in B(0, \sqrt{\frac{b_\pi}{a_\pi}})$, thus the Lipschitz assumption implies

$$\|\nabla V_\pi(z)\|^2 \leq 2L_\pi^2 \left(\|z\|^2 + \frac{b_\pi}{a_\pi}\right). \tag{26}$$

The same properties follow for μ_0.

Using the above and Chehab et al. (2025, Lemma 15) we have

$$\begin{aligned}
&\int \mu_t \left[\Delta V_0 - \Delta V_\pi - \langle \nabla V_\pi, \nabla V_0 - \nabla V_\pi\rangle\right] \\
&\qquad \leq d(L_{\mu_0} - c_\pi) + 2L_\pi^2 \left(\mathbb{E}_{\mu_t}[\|x\|^2] + \frac{b_\pi}{a_\pi}\right) + \int \mu_t(\|\nabla V_\pi\|^2 \vee \|\nabla V_0\|^2) \\
&\qquad \leq d(L_{\mu_0} - c_\pi) + 4(L_\pi^2 \vee L_{\mu_0}^2)\left(\mathbb{E}_{\mu_t}[\|x\|^2] + \left(\frac{b_\pi}{a_\pi} \vee \frac{b_0}{a_0}\right)\right) \\
&\qquad \leq d(L_{\mu_0} - c_\pi) + 4(L_\pi^2 \vee L_{\mu_0}^2)\left(\mathbb{E}_{\mu_0}[\|x\|^2] \vee \frac{d + b_{\mu_0} + b_\pi}{a_{\mu_0} \wedge a_\pi} + \left(\frac{b_\pi}{a_\pi} \vee \frac{b_0}{a_0}\right)\right) := A.
\end{aligned}$$

Using Grönwall's Lemma we obtain

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

A.2 Tempered Fisher Rao PDE

Consider the FR PDE

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

where $\pi_t \propto \pi^{\lambda_t}\mu_0^{1-\lambda_t}$ and $\lambda_t : \mathbb{R}_+ \to [0, 1]$, λ_t is non-decreasing in t and weakly differentiable. Using an argument similar to that of Chen et al. (2023, Appendix B.1) we can show that this PDE reduces the Kullback–Leibler divergence w.r.t. π.

The normalising constant of μ_t satisfies

$$\begin{aligned}
Z_t &:= \int \mu_0(x)^{1-\int_0^t e^{s-t}\lambda_s ds}\pi(x)^{\int_0^t e^{s-t}\lambda_s ds}dx \\
&= \int \pi(x)\left(\frac{\pi(x)}{\mu_0(x)}\right)^{1-\int_0^t e^{s-t}\lambda_s ds} dx \\
&\geq \int \pi(x)\exp\left(-M(1 + \|x\|^2)(1 - \int_0^t e^{s-t}\lambda_s ds)\right)dx \\
&\geq \exp\left(\int \pi(x)\left(-M(1 + \|x\|^2)(1 - \int_0^t e^{s-t}\lambda_s ds)\right)dx\right) \\
&\geq \exp\left(-M(1 + B)(1 - \int_0^t e^{s-t}\lambda_s ds)\right),
\end{aligned}$$

where we used equation 6 and Jensen's inequality.

Since $\int_0^t e^{s-t}\lambda_s ds < 1$ we also have

$$\begin{aligned}\frac{\mu_t(x)}{\pi(x)} &= \frac{1}{Z_t}\frac{\mu_0(x)^{1-\int_0^t e^{s-t}\lambda_s ds}\pi(x)^{\int_0^t e^{s-t}\lambda_s ds}}{\pi(x)} \\ &= \frac{1}{Z_t}\left(\frac{\mu_0(x)}{\pi(x)}\right)^{1-\int_0^t e^{s-t}\lambda_s ds} \\ &\leq \exp\left(M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds) + M(1+\|x\|^2)(1-\int_0^t e^{s-t}\lambda_s ds)\right)\end{aligned}$$

from which we obtain the result

$$\begin{aligned}D_{\mathrm{KL}}(\mu_t||\pi) &\leq M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds) + M(1+\int \mu_t(x)\|x\|^2 dx)(1-\int_0^t e^{s-t}\lambda_s ds) \\ &\leq M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds) + M(1+B\exp\left(M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds)\right))(1-\int_0^t e^{s-t}\lambda_s ds)\end{aligned}$$

using Proposition 1.

Proposition 1. *Assume that* $\int \|x\|^2\mu_0(x)dx \leq B$, $\int \|x\|^2\pi(x)dx \leq B$ *and equation 6, then*

$$\int \mu_t(x)\|x\|^2 dx \leq B\exp\left(M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds)\right).$$

Proof. Using the expression for μ_t in equation 5 and Hölder inequality with $p = (1-\int_0^t e^{s-t}\lambda_s ds)^{-1}$ and $q = (\int_0^t e^{s-t}\lambda_s ds)^{-1}$ gives

$$\begin{aligned}\int \mu_t(x)\|x\|^2 dx &= \frac{1}{Z_t}\int (\mu_0(x)\|x\|^2)^{1-\int_0^t e^{s-t}\lambda_s ds}(\pi(x)\|x\|^2)^{\int_0^t e^{s-t}\lambda_s ds} dx \\ &\leq \frac{1}{Z_t}\left(\int \mu_0(x)\|x\|^2 dx\right)^{1-\int_0^t e^{s-t}\lambda_s ds}\left(\int \pi(x)\|x\|^2 dx\right)^{\int_0^t e^{s-t}\lambda_s ds} \\ &\leq B\exp\left(M(1+B)(1-\int_0^t e^{s-t}\lambda_s ds)\right).\end{aligned}$$

□

B Analytic solutions: Multivariate Gaussian case

This section focuses on comparing the evolution of the PDEs introduce in this section for the special case of transporting a Gaussian $\mu_0(x) = \mathcal{N}(x; m_0, \Sigma_0)$ to target Gaussian $\pi(x) = \mathcal{N}(x; m_\pi, \Sigma_\pi)$. In this case, we can obtain nice expressions for the moments along the flow described by each PDE.

In particular, observing that both the tempered W flow and the tempered FR flow preserve gaussianity, we have that $\mu_t(x) = \mathcal{N}(x; m_t, \Sigma_t)$ for each time $t \geq 0$.

We detail here the results needed to obtain the ODEs describing the tempered and standard PDEs for the special case of transporting a Gaussian $\mu_0(x) = \mathcal{N}(x; m_0, \Sigma_0)$ to target Gaussian $\pi(x) = \mathcal{N}(x; m_\pi, \Sigma_\pi)$.

Wasserstein Flow (Langevin) The PDE in this case is the Fokker-Planck equation for the overdamped Langevin SDE

$$\begin{aligned}dX_t &= -\nabla V_\pi(X_t)dt + \sqrt{2}dB_t \\ &= -\Sigma_\pi^{-1}(X_t - m_\pi)dt + \sqrt{2}dB_t.\end{aligned}$$

Since the potential takes the form $V_\pi(x) = \frac{1}{2}(x - m_\pi)^\top \Sigma_\pi^{-1}(x - m_\pi)$, the moment equations are standard as this essentially boils down to an Ornstein-Uhlenbeck process. They are given by

$$\begin{aligned}\dot{m}_t &= -\Sigma_\pi^{-1}(m_t - m_\pi)\\ \dot{\Sigma}_t &= -\Sigma_\pi^{-1}\Sigma_t - \Sigma_t\Sigma_\pi^{-1} + 2I\end{aligned}$$

and the explicit solution can be found as

$$\begin{aligned}m_t &= e^{-t\Sigma_\pi^{-1}} m_0 + (I - e^{-t\Sigma_\pi^{-1}})m_\pi\\ \Sigma_t &= e^{-t\Sigma_\pi^{-1}}\left(2\int_0^t e^{s\Sigma_\pi^{-1}}(e^{s\Sigma_\pi^{-1}})^\top ds + \Sigma_0\right)(e^{-t\Sigma_\pi^{-1}})^\top.\end{aligned}$$

In the 1D case, we can easily evaluate the expression for Σ_t as

$$\Sigma_t = e^{-2t\Sigma_\pi^{-1}}[\Sigma_0 + \Sigma_\pi(e^{2t\Sigma_\pi^{-1}} - 1)].$$

Fisher–Rao flow In the Gaussian case, the ODEs for the mean and covariance respectively are given by

$$\begin{aligned}\dot{m}_t &= -\Sigma_t\Sigma_\pi^{-1}(m_t - m_\pi)\\ \dot{\Sigma}_t &= -\Sigma_t\Sigma_\pi^{-1}\Sigma_t + \Sigma_t.\end{aligned}$$

The explicit solutions in this case are known (see e.g. Chen et al. (2023, Lemma E.3)),

$$\begin{aligned}m_t &= m_\pi + e^{-t}\left((1 - e^{-t})\Sigma_\pi^{-1} + e^{-t}\Sigma_0^{-1}\right)^{-1}\Sigma_0^{-1}(m_0 - m_\pi)\\ &= m_\pi + e^{-t}\Sigma_t\Sigma_0^{-1}(m_0 - m_\pi)\\ \Sigma_t^{-1} &= \Sigma_\pi^{-1} + e^{-t}(\Sigma_0^{-1} - \Sigma_\pi^{-1}).\end{aligned}$$

Tempered Wasserstein flow (Tempered Langevin; (Chehab et al., 2025)) Recall the diffusion process characterising Tempered Langevin is given by

$$dX_t = -[(1 - \lambda_t)\nabla_x V_0 + \lambda_t\nabla_x V_\pi]dt + \sqrt{2}dB_t$$

where $\lambda_t : \mathbb{R}_+ \to [0, 1)$ is a non-decreasing function of t. The corresponding evolution PDE given by

$$\partial_t\mu_t(x) = \nabla\cdot\left(\mu_t(x)\nabla\log\left(\frac{\mu_t(x)}{\pi_t(x)}\right)\right) = \nabla\cdot\left(\mu_t(x)\nabla\log\left(\frac{\mu_t(x)}{\pi(x)}\right)^{\lambda_t}\right) + \nabla\cdot\left(\mu_t(x)\nabla\log\left(\frac{\mu_t(x)}{\mu_0(x)}\right)^{1-\lambda_t}\right)$$

where $\pi_t \propto \pi^{\lambda_t}\mu_0^{1-\lambda_t}$. The second equality shows that it can be seen as a sum of two Wasserstein parts with temperatures λ_t and $1 - \lambda_t$.

Note that π_t is a product of 2 Gaussian densities $\mathcal{N}\left(m_\pi, \frac{\Sigma_\pi}{\lambda_t}\right)$ and $\mathcal{N}\left(m_0, \frac{\Sigma_0}{1-\lambda_t}\right)$ therefore $\pi_t \propto \mathcal{N}(a_t, P_t)$ where

$$\begin{aligned}a_t &= \left(m_0\frac{\Sigma_\pi}{\lambda_t} + m_\pi\frac{\Sigma_0}{1 - \lambda_t}\right)\left(\frac{\Sigma_0}{1 - \lambda_t} + \frac{\Sigma_\pi}{\lambda_t}\right)^{-1}\\ P_t &= \Sigma_\pi(\lambda_t\Sigma_0 + (1 - \lambda_t)\Sigma_\pi)^{-1}\Sigma_0\end{aligned}$$

and the intermediate densities are still Gaussian. Then we have the same structure for the moment ODEs as for infinite time Wasserstein gradient flow, except with m_π, Σ_π replaced by a_t, P_t respectively,

$$\begin{aligned}\dot{m}_t &= -\left(\Sigma_\pi(\lambda_t\Sigma_0 + (1 - \lambda_t)\Sigma_\pi)^{-1}\Sigma_0\right)^{-1}\left(m_t - \left(m_0\frac{\Sigma_\pi}{\lambda_t} + m_\pi\frac{\Sigma_0}{1 - \lambda_t}\right)\left(\frac{\Sigma_0}{1 - \lambda_t} + \frac{\Sigma_\pi}{\lambda_t}\right)^{-1}\right)\\ &= -(\lambda_t\Sigma_\pi^{-1} + (1 - \lambda_t)\Sigma_0^{-1})\left(m_t - ((1 - \lambda_t)m_0\Sigma_\pi + \lambda_t m_\pi\Sigma_0)(\lambda_t\Sigma_0 + (1 - \lambda_t)\Sigma_\pi)^{-1}\right)\\ \dot{\Sigma}_t &= -\left(\Sigma_\pi(\lambda_t\Sigma_0 + (1 - \lambda_t)\Sigma_\pi)^{-1}\Sigma_0\right)^{-1}\Sigma_t - \Sigma_t\left(\Sigma_\pi(\lambda_t\Sigma_0 + (1 - \lambda_t)\Sigma_\pi)^{-1}\Sigma_0\right)^{-1} + 2I\\ &= -(\lambda_t\Sigma_\pi^{-1} + (1 - \lambda_t)\Sigma_0^{-1})\Sigma_t - \Sigma_t(\lambda_t\Sigma_\pi^{-1} + (1 - \lambda_t)\Sigma_0^{-1}) + 2I.\end{aligned}$$

In the 1D case, the explicit solution is given by

$$
\begin{aligned}
m_t &= m_0 e^{-\int_0^t (1-\lambda_s) ds \Sigma_0^{-1} - \int_0^t \lambda_s ds \Sigma_\pi^{-1}} \\
&\quad + e^{-\int_0^t (1-\lambda_s) ds \Sigma_0^{-1} - \int_0^t \lambda_s ds \Sigma_\pi^{-1}} \int_0^t [(1-\lambda_u)\Sigma_0^{-1} m_0 + \lambda_u \Sigma_\pi^{-1} m_\pi] e^{\int_0^u (1-\lambda_s) ds \Sigma_0^{-1} + \int_0^u \lambda_s ds \Sigma_\pi^{-1}} du \\
\Sigma_t &= \Sigma_0 e^{-2\int_0^t (1-\lambda_s) ds \Sigma_0^{-1} - 2\int_0^t \lambda_s ds \Sigma_\pi^{-1}} \\
&\quad + 2e^{-2\int_0^t (1-\lambda_s) ds \Sigma_0^{-1} - 2\int_0^t \lambda_s ds \Sigma_\pi^{-1}} \int_0^t e^{2\int_0^u (1-\lambda_s) ds \Sigma_0^{-1} + 2\int_0^u \lambda_s ds \Sigma_\pi^{-1}} du.
\end{aligned}
$$

Tempered Fisher–Rao flow Recalling that $\pi_t \propto \mathcal{N}(a_t, P_t)$ with a_t, P_t as before and observing that tempered FR PDE can be written as

$$
\begin{aligned}
\partial_t \mu_t &= \mu_t(x) \left(\log \left(\frac{\pi_t(x)}{\mu_t(x)} \right) - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\pi_t}{\mu_t} \right) \right] \right) \\
&= \mu_t(x) \left(\log \left(\frac{\pi(x)}{\mu_t(x)} \right)^{\lambda_t} - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\pi}{\mu_t} \right)^{\lambda_t} \right] \right) + \mu_t(x) \left(\log \left(\frac{\mu_0(x)}{\mu_t(x)} \right)^{1-\lambda_t} - \mathbb{E}_{\mu_t} \left[\log \left(\frac{\mu_0}{\mu_t} \right)^{1-\lambda_t} \right] \right),
\end{aligned}
$$

we see that the tempered FR PDE is the sum of two Fisher–Rao paths with temperatures λ_t and $1 - \lambda_t$.

The ODEs describing the evolution of the moments of μ_t are obtained from those for the standard FR flow replacing m_π, Σ_π with a_t, P_t

$$
\begin{aligned}
\dot{m}_t &= -\Sigma_t(\lambda_t \Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})(m_t - m_\pi) \\
\dot{\Sigma}_t &= -\Sigma_t(\lambda_t \Sigma_\pi^{-1} + (1-\lambda_t)\Sigma_0^{-1})\Sigma_t + \Sigma_t.
\end{aligned}
$$

C Discrete time convergence

C.1 Tempered ULA

To deal with the tempered W flow update we use Step 2 in the proof of Chehab et al. (2025, Theorem 3) to obtain

$$
D_{\mathrm{KL}}(\mu_n || \pi) = D_{\mathrm{KL}}(\mu_n || \pi_n) + \int \mu_n(x) \log \frac{\pi_n(x)}{\pi(x)} dx. \tag{27}
$$

We recall that μ_n is obtained as

$$
\mu_n(x) = (4\pi\gamma_n)^{-d/2} \int \exp\left(-\frac{1}{4\gamma_n} \|x - y - \gamma_n \nabla \log \pi_n(y)\|^2 \right) \mu_{n-1}(y) dy
$$

and using Vempala & Wibisono (2019, Lemma 3) with $\gamma_n \leq c_n/(4L_n^2)$ we obtain

$$
D_{\mathrm{KL}}(\mu_n || \pi_n) \leq e^{-\gamma_n c_n} D_{\mathrm{KL}}(\mu_{n-1} || \pi_n) + 6\gamma_n^2 d L_n^2.
$$

Thus we obtain

$$
\begin{aligned}
D_{\mathrm{KL}}(\mu_n || \pi) &\leq e^{-\gamma_n c_n} D_{\mathrm{KL}}(\mu_{n-1} || \pi_n) + 6\gamma_n^2 d L_n^2 + \int \mu_n(x) \log \frac{\pi_n(x)}{\pi(x)} dx \\
&\leq e^{-\gamma_n c_n} D_{\mathrm{KL}}(\mu_{n-1} || \pi) + 6\gamma_n^2 d L_n^2 + \int [\mu_{n-1}(x) - \mu_n(x)] \log \frac{\pi(x)}{\pi_n(x)} dx \\
&= e^{-\gamma_n c_n} D_{\mathrm{KL}}(\mu_{n-1} || \pi) + 6\gamma_n^2 d L_n^2 + (1-\lambda_n) \int [\mu_{n-1}(x) - \mu_n(x)] \log \frac{\pi(x)}{\mu_0(x)} dx,
\end{aligned}
$$

where we used the fact that $D_{\mathrm{KL}}(\mu_{n-1}||\pi_n) = D_{\mathrm{KL}}(\mu_{n-1}||\pi) + \int \mu_{n-1} \log {}^{\pi}/_{\pi_n}$ and $e^{-\gamma_n c_n} \leq 1$ and that $\pi_n = \mu_0^{1-\lambda_n}\pi^{\lambda_n}/\mathcal{Z}_n$.

Let us focus on

$$\int [\mu_{n-1}(x) - \mu_n(x)] \log \frac{\pi(x)}{\mu_0(x)} dx = \int \mu_{n-1}(x)(V_0(x) - V_\pi(x))dx + \int \mu_n(x)(V_\pi(x) - V_0(x))dx$$
$$\leq \int \mu_{n-1}(x)[V_0(x) - \inf_{x\in\mathbb{R}^d} V_0(x)]dx + \int \mu_n(x)[V_\pi(x) - \inf_{x\in\mathbb{R}^d} V_\pi(x)]dx.$$

Using Chehab et al. (2025, Proposition 14) under equation 24, we can bound

$$\int [\mu_{n-1}(x) - \mu_n(x)] \log \frac{\pi(x)}{\mu_0(x)} dx \leq 2L_0 \frac{b_0}{a_0} + 2L_\pi \frac{b_\pi}{a_\pi} + 2L_0 \int \mu_{n-1}(x)\|x\|^2 dx + 2L_\pi \int \mu_n(x)\|x\|^2 dx$$
$$\leq 2L_0 \frac{b_0}{a_0} + 2L_\pi \frac{b_\pi}{a_\pi} + 2(L_0 + L_\pi) \max\left(\mathbb{E}_{\mu_0}[\|x\|^2], \frac{2(3(b_0 + b_\pi)/2 + d)}{a_0 \wedge a_\pi \wedge 1}\right) := A'$$

using Chehab et al. (2025, Lemma 16) to bound the second moment of μ_n for all $n \geq 0$ since $\gamma_n \leq \min(1, \frac{a_\pi \wedge a_0}{2(L_\pi + L_0)^2})$ for all $n \geq 0$.

Using induction we then obtain

$$D_{\mathrm{KL}}(\mu_n||\pi) \leq D_{\mathrm{KL}}(\mu_0||\pi)e^{-\sum_{k=1}^n \gamma_k c_k} + \sum_{k=0}^{n-1} (6\gamma_k^2 d L_k^2 + (1-\lambda_k)A')e^{-\sum_{i=1}^k \gamma_i c_i}.$$

D Auxiliary results for Section 3

D.1 First variation of $\mathcal{F}$

We recall the definition of first variation of a functional $\mathcal{F} : \mathsf{E} \to \mathbb{R}^+$ on E. We denote by E^* the dual of E and for any $z \in \mathsf{E}$ and $z^\star \in \mathsf{E}^*$ we denote the dot product by $\langle z^\star, z\rangle$.

Definition 1 (Derivative)**.** *If it exists, the derivative of $\mathcal{F}$ at z_1 is the function $\delta\mathcal{F}(z_1) : \mathsf{E}^\star \to \mathbb{R}$ s.t. for any $z_2 \in \mathsf{E}$, with $\xi = z_2 - z_1$:*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}(\mathcal{F}(z_1 + \epsilon\xi) - \mathcal{F}(z_1)) = \langle \delta\mathcal{F}(z_1), \xi\rangle,$$

and is defined uniquely up to an additive constant.

Using this definition for $\mathcal{F}(\mu, \lambda) = D_{\mathrm{KL}}(\mu||\pi_\lambda)$ we have for $z_i = (\mu_i, \lambda_i)$, $i = 1, 2$, and $\xi = z_2 - z_1 = (\xi_\lambda, \xi_\mu)$,

$$\mathcal{F}(z_1 + \epsilon\xi) - \mathcal{F}(z_1) = \int \log\left([\mu_1 + \epsilon\xi_\mu](x)\right)[\mu_1 + \epsilon\xi_\mu](x)dx - \int \log \pi_{\lambda_1+\epsilon\xi_\lambda}(x)[\mu_1 + \epsilon\xi_\mu](x)dx +$$
$$+ \int \log \pi_{\lambda_1}(x)\mu_1(x)dx - \int \log\left(\mu_1(x)\right)\mu_1(x)dx$$
$$= -\int [\log \pi_{\lambda_1+\epsilon\xi_\lambda}(x) - \log \pi_{\lambda_1}(x)]\mu_1(x)dx - \epsilon \int \log \pi_{\lambda_1+\epsilon\xi_\lambda}(x)\xi_\mu(x)dx$$
$$+ \int \log\left(1 + \epsilon\frac{\xi_\mu(x)}{\mu_1(x)}\right)\mu_1(x)dx + \epsilon \int \log\left([\mu_1 + \epsilon\xi_\mu](x)\right)\xi_\mu(x)dx.$$

We then have that

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}(\mathcal{F}(z_1 + \epsilon\xi) - \mathcal{F}(z_1)) = -\xi_\lambda \int \partial_\lambda \log \pi_\lambda(x)|_{\lambda=\lambda_1} \mu_1(x)dx - \int \log \pi_{\lambda_1}(x)\xi_\mu(x)dx$$
$$+ \int \frac{\xi_\mu(x)}{\mu_1(x)}\mu_1(x)dx + \int \log\left(\mu_1(x)\right)\xi_\mu(x)dx,$$

where the equality follows from the Taylor expansion of the logarithm as $\epsilon \to 0$. Therefore, we can write

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(z_1 + \epsilon\xi) - \mathcal{F}(z_1)) = \left\langle \begin{matrix} -\int \partial_\lambda \log \pi_\lambda(x)|_{\lambda=\lambda_1} \mu_1(x)dx \\ \log \frac{\mu_1(x)}{\pi_{\lambda_1}(x)} + 1 \end{matrix}, \begin{matrix} \xi_\lambda \\ \xi_\mu \end{matrix} \right\rangle.$$

The result follows since $\delta\mathcal{F}$ is defined up to additive constants.