A Connection between Tempering and Entropic Mirror Descent

Nicolas Chopin [1] **Francesca Crucinio** [1] **Anna Korba** [1]

Abstract

This paper explores the connections between tempering (for Sequential Monte Carlo; SMC) and entropic mirror descent to sample from a target probability distribution whose unnormalized density is known. We establish that tempering SMC corresponds to entropic mirror descent applied to the reverse Kullback-Leibler (KL) divergence and obtain convergence rates for the tempering iterates. Our result motivates the tempering iterates from an optimization point of view, showing that tempering can be seen as a descent scheme of the KL divergence with respect to the Fisher-Rao geometry, in contrast to Langevin dynamics that perform descent of the KL with respect to the Wasserstein-2 geometry. We exploit the connection between tempering and mirror descent iterates to justify common practices in SMC and derive adaptive tempering rules that improve over other alternative benchmarks in the literature.

1. Introduction

Sampling from a target probability distribution whose density is known up to a normalization constant is a fundamental task in computational statistics and machine learning. It can be naturally formulated as optimizing a functional measuring the dissimilarity to the target probability distribution, typically the Kullback-Leibler (KL) divergence. From there, it is natural to consider optimization schemes over the space of probability distributions, to design a sequence of distributions approximating the target one. Depending on the chosen geometry over the search space and the time discretization, one may obtain different schemes.

For instance, one possible framework is to restrict the search space to the Wasserstein space, i.e. probability distributions with bounded second moments equipped with the Wasserstein-2 distance (Ambrosio et al., 2008). The latter is equipped with a rich Riemannian structure (Otto and Villani, 2000), which makes it possible to define Wasserstein-2 gradient flows, i.e. paths of distributions decreasing the objective functional of steepest descent according to this metric. It is well-known that the Wasserstein gradient flow of the KL can be implemented by a Langevin diffusion on the ambient space (Jordan et al., 1998) and easily discretized in time, resulting for instance in the Langevin Monte Carlo (or Unadjusted Langevin) algorithm (Roberts and Tweedie, 1996). The latter is one of the most famous Markov Chain Monte Carlo (MCMC) algorithms - maybe the most canonical - that generate Markov chains in the ambient space, whose law approximates the target distribution for a large time horizon. It is known to converge fast when the target distribution has a smooth and strongly convex potential (Durmus et al., 2019), or is satisfies a relaxed log-Sobolev assumption (Vempala and Wibisono, 2019). Alternative time discretizations of the KL Wasserstein gradient flow (Salim et al., 2020; Mou et al., 2021) or its gradient flow with respect to similar optimal transport geometries have been considered in the literature to propose alternative algorithms (Liu, 2017; Garbuno-Inigo et al., 2020), but their convergence also depends strongly on the log-concavity of the target.

Another possible framework is to cast the space of probability distributions as a subset of a normed space of measures (such as L^2), and to consider the duality of measures with continuous functions and the mirror descent algorithm that relies on Bregman divergences geometry, as recently considered in Ying (2020); Chizat (2022); Aubin-Frankowski et al. (2022). While both frameworks yield optimization algorithms on measure spaces, the geometries and algorithms are very different (in particular notions of gradients and convexity). Mirror descent produces multiplicative ("vertical") updates on measures allowing for change of mass, while Wasserstein flows corresponds to displacement of fixed mass particles supporting the measures ("horizontal" updates). Moreover, as recently highlighted in Aubin-Frankowski et al. (2022), the (reverse) KL as an objective loss for sampling is actually strongly convex and smooth *whatever the target* π in a mirror descent geometry induced by the KL as a Bregman divergence. In contrast, it is known that

[1]ENSAE, CREST, Institut Polytechnique de Paris. Correspondence to: Nicolas Chopin <Nicolas.Chopin@ensae.fr>, Francesca Crucinio <francesca_romana.crucinio@kcl.ac.uk>, Anna Korba <anna.korba@ensae.fr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

the KL as an optimization objective is not smooth in the Wasserstein geometry (Wibisono, 2018), and as we have said earlier it enjoys convexity properties only if the target distributions does as well. Above all, the latter scheme, namely *entropic mirror descent* on the KL yields a sequence of distributions that takes the simple form of a geometric mixture between an initial distribution and the target, a well-known sequence referred to as *tempering* (or *annealing*) in the Monte Carlo literature (Neal, 2001). Interestingly, entropic mirror descent on an objective functional can be seen as an Euler discretization of its Fisher Rao gradient flow (Domingo-Enrich and Pooladian, 2023); gradient flows in this geometry have recently attracted a lot of interest thanks to their nice theoretical properties (Chen et al., 2023; Yan et al., 2023).

Algorithms approximating the tempering sequence offer an alternative to Langevin-based MCMC methods, and are often employed when the latter suffer from poor mixing (Syed et al., 2022) or when estimates of the normalizing constant are needed (Gelman and Meng, 1998). A number of algorithms have been proposed to approximate the tempering sequence, including sequential Monte Carlo (SMC; Del Moral et al. (2006)), annealed importance sampling (AIS; i.e. an SMC sampler in which no resampling occurs Neal (2001)), and parallel tempering (PT; Geyer (1991)). Independently, a number of schemes aiming at directly approximating the entropic mirror descent iterates on the KL have also been proposed (Dai et al., 2016; Korba and Portier, 2022).

Choosing the right scheduling of temperatures for the sequence of tempered targets (or equivalently the step-sizes as we will explain in more detail in this paper), is critical in practice. Adaptive selection of the sequence of temperatures is an active area of research in the AIS literature; however, many of these strategies are intractable (Gelman and Meng, 1998), costly (Kiwaki, 2015), limited to exponential families (Grosse et al., 2013), or numerically unstable (Goshtasbpour et al., 2023) as we show in our experiments. In the SMC literature, the sequence of temperatures is normally chosen adaptively using the effective sample size, a proxy for the variance of the importance sampling weights (Jasra et al., 2011). Adaptive strategies are widely used in practice but theoretical studies on how to select the tempering iterates are limited to specific target distributions (see Beskos et al. (2014) for i.i.d. targets and Chopin and Papaspiliopoulos (2020, Proposition 17.2), Dai et al. (2022, Section 3.3) for Gaussian targets).

In this paper, we investigate the links between tempering and mirror descent and show that algorithms which sample from the tempering sequence (such as SMC) can be seen as numerical approximations to entropic mirror descent applied to the KL divergence, i.e. a time-discretization of the KL gradient flow in the Fisher Rao geometry. We thus establish a parallel result to that of Jordan et al. (1998); Wibisono (2018) which shows that algorithms based on the Langevin diffusion can be seen as numerical approximations of gradient flow of the KL in the Wasserstein-2 geometry.

We adapt the proof of convergence of mirror descent in the space of measures of Aubin-Frankowski et al. (2022, Theorem 4) to the case of varying step sizes and obtain a convergence rate for the tempering iterates. From this optimization point of view, we also justify the popular adaptive strategy that identifies the tempering sequence by ensuring that the (KL, Bregman) divergence between two consecutive distributions in the tempered sequence is small and constant. We show that for a generic target distribution, this tempering sequence obeys a differential equation, that can be solved easily analytically in some simple cases that we highlight, or by a simple numerical approximation based on particles in general cases.

The paper is organized as follows. Section 2 provides the relevant background on mirror descent on the space of measures. Section 3 details the connection between tempering and entropic mirror descent and its consequence on designing tempering schedules. Section 4 discusses different strategies that were employed in the literature to approximate entropic mirror descent and their pros and cons. In Section 5 we connect our results with relevant works in the SMC/AIS literature.

2. Mirror descent on measures

In this section, we recall the main steps to derive the mirror descent algorithm on the space of measures (see Appendix C for more details). The reader may refer to Aubin-Frankowski et al. (2022) for a detailed introduction.

Notations. Fix a vector space of (signed) measures $\mathcal{M}(\mathbb{R}^d)$. Let $\mathcal{M}^*(\mathbb{R}^d)$ the dual of $\mathcal{M}(\mathbb{R}^d)$. For $\mu \in \mathcal{M}(\mathbb{R}^d)$ and $f \in \mathcal{M}^*(\mathbb{R}^d)$, we denote $\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(x)\mu(dx)$. We denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on $\mathbb{R}^d$. The Kullback-Leibler divergence is defined as follows: for $\nu, \mu \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{KL}(\nu|\mu) = \int \log(d\nu/d\mu)d\nu$ if ν is absolutely continuous w.r.t. μ with Radon-Nikodym density $d\nu/d\mu$, and $+\infty$ else.

2.1. Background on Mirror Descent

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

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

Mirror descent is a first-order optimization scheme relying on the knowledge of the derivatives of the objective functional, and a geometry on the search space induced by

Bregman divergences. These two notions are introduced in the following definitions.

Definition 1. *If it exists, the first variation of $\mathcal{F}$ at ν is the function $\nabla\mathcal{F}(\nu) : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu \in \mathcal{P}(\mathbb{R}^d)$, with $\xi = \mu - \nu$:*

$$\lim_{\epsilon \to 0} {}^1\!/\!_\epsilon (\mathcal{F}(\nu + \epsilon\xi) - \mathcal{F}(\nu)) = \langle \nabla\mathcal{F}(\nu), \xi \rangle \tag{1}$$

and is defined uniquely up to an additive constant.

Definition 2. *Let $\phi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ a convex functional on $\mathcal{P}(\mathbb{R}^d)$. The ϕ-Bregman divergence is defined for any $\nu, \mu \in \mathcal{P}(\mathbb{R}^d)$ by:*

$$B_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - \langle \nabla\phi(\mu), \nu - \mu \rangle \tag{2}$$

where $\nabla\phi(\mu)$ is the first variation of ϕ at μ.

Consider a sequence of step-sizes $(\gamma_n)_{n\geq 0}$. Starting from an initial $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, one can generate a sequence $(\mu_n)_{n\in\mathbb{N}}$

$$\mu_{n+1} = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \{\mathcal{F}(\mu_n) + \langle \nabla\mathcal{F}(\mu_n), \mu - \mu_n \rangle + (\gamma_{n+1})^{-1} B_\phi(\mu|\mu_n)\}. \tag{3}$$

The first variation of ϕ, denoted $\nabla\phi : \mathcal{P}(\mathbb{R}^d) \to \mathcal{C}(\mathbb{R}^d)$, maps an element of the primal (a distribution) to an element of the dual (a function). In particular, writing the first order conditions of (3) we obtain the dual iteration

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\gamma_{n+1}\nabla\mathcal{F}(\mu_n). \tag{4}$$

The scheme (3)-(4) is referred to as mirror descent (Beck and Teboulle, 2003). It has been shown recently in Aubin-Frankowski et al. (2022), that the mirror descent scheme converges linearly as soon as there exists $0 \leq l \leq L$ such that $\mathcal{F}$ is relatively l-strongly convex and L-smooth with respect to ϕ, a condition that can be written as $lB_\phi(\nu|\mu) \leq B_\mathcal{F}(\nu|\mu) \leq LB_\phi(\nu|\mu)$, for constant stepsizes smaller than ${}^1\!/\!_L$; extending the results of Lu et al. (2018) to the infinite-dimensional setting of optimization over measures. In particular it applies to the case where both the objective and Bregman divergence are chosen as the KL.

2.2. Entropic mirror descent on the KL

Consider the negative entropy functional:

$$\phi : \mu \mapsto \int \log(\mu(x)) d\mu(x) \tag{5}$$

where μ also denotes its density w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Since the first variation of ϕ at μ writes $\nabla\phi(\mu) = \log(\mu)$, one gets from (2) that $B_\phi(\nu|\mu) = \mathrm{KL}(\nu|\mu)$, and choosing ϕ as (5) yields the following multiplicative update named *entropic mirror descent*:

$$\mu_{n+1} \propto \mu_n e^{-\gamma_{n+1}\nabla\mathcal{F}(\mu_n)} \tag{6}$$

by exponentiating (4). Notice that the latter scheme is an Euler discretization of the Fisher-Rao gradient flow of the functional $\mathcal{F}$, as noticed in Domingo-Enrich and Pooladian (2023) (see Appendix D for details).

Moreover, if $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ (the reverse KL with respect to π), $\nabla\mathcal{F}(\mu) = \log({}^\mu\!/\!_\pi)$ and we obtain *entropic mirror descent* on the KL iterates:

$$\mu_{n+1} \propto \mu_n^{(1-\gamma_{n+1})}\pi^{\gamma_{n+1}}. \tag{7}$$

Since $\mathcal{F}$ is 1-strongly convex and 1-smooth with respect to ϕ since $B_\mathcal{F} = B_\phi$ (i.e. $l = L = 1$), as soon as one uses step-sizes $\gamma_n < 1$, the KL objective decreases at each step of the scheme (7), and converges at a linear rate as stated in the Proposition below.

Proposition 1. *Let $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ an initial distribution. Entropic mirror descent iterates on $\mathcal{F} = \mathrm{KL}(\cdot|\pi)$ as defined in (7) converge at a rate:*

$$\mathrm{KL}(\mu_n|\pi) \leq \frac{C_n}{\gamma_1}\mathrm{KL}(\pi|\mu_0);\ C_n^{-1} = \sum_{k=1}^{n}\prod_{i=1}^{k}\frac{\gamma_k/\gamma_1}{1-\gamma_i}. \tag{8}$$

where $(\gamma_k)_{k=1}^n$ is the sequence of step-sizes. In particular, a simple induction argument shows that $C_n \leq \prod_{k=1}^n(1-\gamma_k) \to 0$ as $n \to \infty$ when $\gamma_n \leq 1$ for all $n \geq 1$. Hence, the mirror descent iterates (7) satisfy

$$\mathrm{KL}(\mu_n|\pi) \leq (\gamma_1)^{-1}\prod_{k=1}^{n}(1-\gamma_k)\,\mathrm{KL}(\pi|\mu_0). \tag{9}$$

The proof of Proposition 1 is given in Appendix A. It extends the result of Aubin-Frankowski et al. (2022, Theorem 4) to the case of varying step-sizes, $\gamma_n \neq \gamma$ for all $n \geq 1$. We derived our result by carefully adapting the proof of Aubin-Frankowski et al. (2022, Theorem 4) or Lu et al. (2018, Theorem 3.1); the extension is non-trivial and involves verifying a recursion that is the same as the one in the latter references for constant or decreasing step sizes (as detailed in Appendix A.1) or a different one for general step-sizes (as detailed in Appendix A.2). We consider the two cases separately, as this allows us to obtain sharper rates.

Proposition 1 shows that if $\mathrm{KL}(\pi|\mu_0) < \infty$ and the step sizes are smaller than 1 (the inverse of the smoothness constant $L = 1$), $C_n^{-1} \to \infty$, and entropic mirror descent on the KL converges to the target distribution. We note that Korba and Portier (2022, Lemma 2) show a similar result on the total variation[1]. We also note that our (discrete-time) rate is coherent with the convergence rate of its continuous-time counterpart, i.e. Fisher-Rao dynamics for the KL (Lu et al., 2023, Theorem 2.4), that is known to converge exponentially fast under a warm-start assumption on the support

[1] notice that Pinsker's inequality combined with our result (9) recover their rate on the TV.

of the initial distribution with respect to the target. Finally, if the sequence of $(\gamma_n)_{n\geq 0}$ is fixed to γ constant, mirror descent converges at a linear rate proportional to $(1-\gamma)^n$, as already shown in Lu et al. (2018, Eq. (27)).

3. A connection between Mirror Descent and Tempering

We now turn to the connection between entropic mirror descent and tempering, that, to the best of our knowledge, we are the first to to highlight and exploit (see Domingo-Enrich and Pooladian (2023) for a similar connection in continous time). In particular we will show that the tempering schedule is deeply connected to optimization/step-sizes dynamics of the corresponding entropic mirror descent scheme.

In the Monte Carlo literature, it is common to consider the following tempering (or annealing) sequence (Gelman and Meng, 1998; Neal, 2001)

$$\mu_{n+1} \propto \mu_0^{1-\lambda_{n+1}} \pi^{\lambda_{n+1}}, \tag{10}$$

where $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = 1$, to sample from a target distribution π. There is a correspondence between (10) and (7) if

$$\lambda_n = 1 - \prod_{k=1}^{n} (1-\gamma_k) \tag{11}$$

which by induction yields $\gamma_1 = \lambda_1$, $\gamma_n = (\lambda_n - \lambda_{n-1})/1-\lambda_{n-1}$ for $1 \leq n < T$ and $\gamma_T = \lambda_T = 1$. Notice that reversely, if we have a sequence γ_n defined as $\gamma_n = (\lambda_n - \lambda_{n-1})/1-\lambda_{n-1}$, as soon as the λ's are in $(0,1)$, $\gamma_n < 1$, guaranteeing descent of the KL objective at each step.

In the tempering sequence (10), $\lambda_T = 1$ to ensure that we are targeting the correct distribution π. In the case of the mirror descent iterates (7) the convergence to π is in the limit $n \to \infty$. We can thus interpret (10) as performing $T-1$ mirror descent steps towards π and then one final bridging step to reach π. Hence, it is interesting to look at the speed of convergence of the iterates (7) to gain some intuition on the number of bridging distributions μ_n necessary to get close enough to π to guarantee that the final step (corresponding to $\lambda_T = 1$) is stable. In this case, combining and (9) and (11) we get that

$$\mathrm{KL}(\mu_n|\pi) \leq (\lambda_1)^{-1}(1-\lambda_n)\,\mathrm{KL}(\pi|\mu_0), \tag{12}$$

which approaches 0 as $\lambda_n \to 1$, and gives an explicit rate of convergence of the sequence (10). Provided one can obtain an approximation of $\mathrm{KL}(\pi|\mu_0)$, we can infer the value of λ_n necessary to guarantee that the $n-$th tempering iterate is sufficiently close to π. Later in this section, we derive several examples.

3.1. A principled strategy for tempering

As the speed of convergence of the mirror descent iterates depends on the sequence $(\lambda_n)_{n\geq 1}$, we now discuss relevant strategies to select temperatures, in the light of the optimization scheme.

Notice that (10) admits an exponential family representation (Brekelmans et al., 2020; Syed et al., 2021)

$$\mu_{n+1}(x) \equiv \mu_{\lambda_{n+1}}(x) \propto \mu_0 \exp\{\lambda_{n+1} s(x)\} \tag{13}$$

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

A popular strategy in the SMC/AIS literature to identify the sequence $(\lambda_n)_{n=0}^T$ is to fix $\lambda_0 = 0$ and then select λ_n iteratively, ensuring that the χ^2 divergence between successive distributions is constant and sufficiently small, e.g. setting $\chi^2(\mu_{n-1}|\mu_n) = \beta$ for some small value of β (see Jasra et al. (2011) for χ^2 in SMC, and more recently Goshtasbpour et al. (2023) for α-divergences in AIS). This quantity is related to the variance of the importance weights and ensures that this variance remains low.

The following Proposition, whose proof can be found in Appendix B, shows that, up to higher order terms, the χ^2-divergence can be replaced by any $f-$divergence whose f is twice differentiable (see also Amari (2016, Section 3.4)), in particular the KL divergence. Let $D_f(\lambda'|\lambda) := \int \mu_\lambda f(\mu_{\lambda'}/\mu_\lambda)$ be the $f-$divergence of $\mu_{\lambda'}$ relative to μ_λ.

Proposition 2. *Provided f is twice differentiable, one has:*

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

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

This proposition applies in particular to the KL divergence ($f(x) = x\log x$, $f''(1) = 1$), the reverse KL ($f(x) = -\log x$, $f''(1) = 1$), all $\alpha-$divergences ($f''(1) = 1$), the χ^2-divergence ($f(x) = (x-1)^2$, $f''(1) = 2$), hence fixing the χ^2-divergence constant or the KL between consecutive iterates only differs by a multiplicative factor (resp. β or $\beta/2$).

The tempering strategy previously described can be justified by looking at the convergence of the corresponding entropic mirror descent scheme on the KL (7) where both $\mathcal{F}$ and B_ϕ are chosen as the KL divergence. Indeed, as shown in Eq. (24) in Appendix A, the (Bregman) divergence between iterates $B_\phi(\mu_{n-1}|\mu_n) = \mathrm{KL}(\mu_{n-1}|\mu_n)$ provides a lower bound on the decay of the objective $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ achieved by one iteration of mirror descent : since $\gamma_n \leq L^{-1} = 1$ for all n, we have

$$\mathrm{KL}(\mu_{n-1}|\pi) - \mathrm{KL}(\mu_n|\pi) \geq \mathrm{KL}(\mu_{n-1}|\mu_n) = \frac{\beta}{2}. \tag{14}$$

Proposition 2 above suggests the following recipe to choose successive λ_n values (see Iba (2001, Eq. (10)) for a similar result in the PT literature):

$$\lambda_n - \lambda_{n-1} = c\mathrm{I}(\lambda_{n-1})^{-1/2} \tag{15}$$

for a certain $c \geq 0$; we recover $\chi^2(\mu_{n-1}|\mu_n) = \beta$ taking $c = \sqrt{\beta}$ (standard practice is to take $\beta = 1$, which is equivalent to ESS $= N/2$, see Section 4). For a model where π and μ_0 correspond to the distribution of d i.i.d. components, it is well known that $\mathrm{I}(\lambda) = d\mathrm{I}_1(\lambda)$ where I_1 denotes the Fisher information corresponding to one component. We automatically get therefore that the number of successive steps to move from $\lambda_0 = 0$ to $\lambda_T = 1$ should be $\mathcal{O}(d^{1/2})$, as already observed by Chopin and Papaspiliopoulos (2020, Proposition 17.2) and Dai et al. (2022, Section 3.3) in the context of SMC (for Gaussian targets) and Atchadé et al. (2011, Section 2.3) in the context of PT. We point out that Proposition 2 is a result about the sequence of distributions (10) and not about a specific class of algorithms like the results of Beskos et al. (2014). See Figure 2 for a numerical experiment illustrating this point.

3.2. Examples of tempering sequences

We now consider the simplified scenario in which π and μ_0 correspond to the distribution of d i.i.d. components. For some examples of proposals μ_0 and targets π the optimal tempering sequence satisfying (15) can be found (at least for large d) analytically. Our aim is to use the correspondence between γ_n and λ_n to obtain the convergence rate of the corresponding mirror descent scheme.

For large d, it makes sense to replace the sequence λ_n by a continuous function $\lambda(t)$, and solve the ODE:

$$\dot{\lambda} = c\mathrm{I}(\lambda)^{-1/2}. \tag{16}$$

As a first simple case, consider targeting a pair of Gaussians with the same mean but different variances. Let $\pi = \mathcal{N}(0, \tau^2 \mathrm{Id})$ starting from $\mu_0 = \mathcal{N}(0, \mathrm{Id})$. In this case $s(x) = x^2(1 - 1/\tau^2)/2 - \log\tau$, and we have $\mu_\lambda = \mathcal{N}\left(0, \left(1 - \lambda + \lambda/\tau^2\right)^{-1} \mathrm{Id}\right)$ and $\mathrm{I}(\lambda) \propto \left(1 - \lambda + \lambda/\tau^2\right)^{-2}$. The corresponding ODE is $\dot{\lambda} = c(1 + \alpha\lambda)$ with $\alpha = 1/\tau^2 - 1$. If $\tau > 1$, then $\alpha < 0$, and the solution is $\lambda(t) = 1 - \exp(\alpha t)$, which behaves likes a *negative exponential*. This corresponds to a constant $\gamma = 1 - \exp(\alpha)$. Conversely, if $\tau < 1$, then $\alpha > 0$, and the solution is $\lambda(t) = \exp(\alpha t) - 1$, which corresponds to *exponential growth*.

We then consider the case in which the variance is the same, but the means are different. Let $\mu_0 = \mathcal{N}(0, \mathrm{Id})$ and $\pi = \mathcal{N}(m, \mathrm{Id})$, so μ_λ is $\mathcal{N}(\lambda m, \mathrm{Id})$, $s(x) = mx - m^2/2$, and $\mathrm{I}(\lambda) = m^2$ is constant. In this case, $\lambda(t) = mt$ grows *linearly*.

For each of the examples of tempering sequences $(\lambda_n)_{n\geq 1}$ exhibited in this subsection, we have seen at the beginning of this section that the tempered sequence converges at a rate $C_n \leq 1 - \lambda_n$. Figure 1 provide some illustrations of the joint evolutions of temperatures $(\lambda_n)_{n\geq 1}$, mirror descent step sizes $(\gamma_n)_{n\geq 1}$ and rate of convergence $(C_n)_{n\geq 1}$ in these three different Gaussian scenarios.

4. Algorithmic approximations

Having identified the connection between the mirror descent iterates (7) and the tempering iterates (10), we now turn to existing (and potentially improvable) algorithms approximations, and identify their connections. Indeed, while (7) is attractive for its nice convergence properties, it is not feasible to run this iteration in practice for several reasons: each iteration depends on the whole densities, and it requires a normalization step.

Notice from (6) that it is natural to approximate entropic mirror descent on $\mathcal{F} = \mathrm{KL}(\cdot|\pi)$ by

$$\mu_{n+1} \propto q_n \exp(-\gamma_n g_n) \tag{17}$$

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 discuss here a common strategy in the Monte Carlo literature to approximate (10) based on importance sampling and show that the exponential update is performed on the importance weights.

Sequential Monte Carlo (SMC) samplers Del Moral et al. (2006) provide a particle approximation of the tempering iterates (10) using clouds of N weighted particles $\{X_n^i, W_n^i\}_{i=1}^N$. The fundamental ingredients of an SMC sampler are the sequence $(\lambda_n)_{n=0}^T$ with $0 = \lambda_0 < \cdots < \lambda_T = 1$, a family of Markov kernels $(M_n)_{n=1}^T$ used to propagate the particles forward in time and a resampling scheme.

For simplicity, we focus here on the case in which the Markov kernels M_n are μ_n-invariant, the resulting SMC algorithm is summarized in Algorithm 1 in Appendix E. At iteration n the weighted particle set $\{X_{n-1}^i, W_{n-1}^i\}_{i=1}^N$ is resampled to obtain the equally weighted particle set $\{\widetilde{X}_{n-1}^i, 1/N\}_{i=1}^N$ and the kernel M_n is applied to propose new particle locations $X_n^i \sim M_n(\cdot, \widetilde{X}_{n-1}^i)$. The weights are proportional to

$$w_n(x) = \frac{\eta_n(x)}{\eta_{n-1}(x)} = \left(\frac{\pi(x)}{\mu_0(x)}\right)^{\lambda_n - \lambda_{n-1}} \tag{18}$$

where $\eta_n := \mu_0^{1-\lambda_n}\pi^{\lambda_n}$ and $\mu_n = \eta_n/Z_n$. Recalling the relationship between the sequence of γ_n and of λ_n, $\gamma_n =$

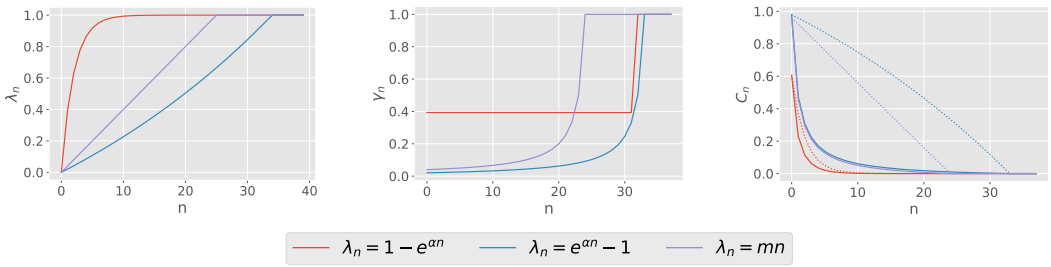

Figure 1: Sequence of $(\lambda_n)_{n\geq 1}$, $(\gamma_n)_{n\geq 1}$ and rate C_n for the negative exponential, positive exponential and linear evolution of $(\lambda_n)_{n\geq 1}$. The dotted lines in the right-most plot show the bound $1 - \lambda_n$ on C_n.

$(\lambda_n - \lambda_{n-1})/(1 - \lambda_{n-1})$, we find that

$$w_n(x) = \left(\frac{\pi(x)}{\eta_{n-1}(x)}\right)^{\gamma_n} \propto \left(\frac{\pi(x)}{\mu_{n-1}(x)}\right)^{\gamma_n}, \qquad (19)$$

where the normalizing constant can be discarded due to the re-normalization, showing that the importance weights used within an SMC sampler approximate the exponential update in (17). The approximation of μ_n provided by SMC is $q_n^{\mathrm{SMC}}(x) = \sum_{i=1}^N W_n^i \delta_{X_n^i}(x)$, where $\delta(\cdot)$ denotes the Dirac's delta function and $W_n^i = w_n(\widetilde{X}_{n-1}^i)$.

Remark 1. *In Appendix E we discuss two alternative strategies to SMC based on importance sampling that directly approximate (6) (Dai et al., 2016; Korba and Portier, 2022). We highlight in particular that MD on measures can be implemented through SMC with a better complexity than the scheme proposed in Dai et al. (2016) (the weights (18) do not depend on the N particle set and can be computed in $\mathcal{O}(1)$ time while those of Dai et al. (2016) depend on the N particle set and require $\mathcal{O}(N)$ cost, see Appendix E).*

In the SMC literature, the tempering schedule $\{\lambda_n\}_{n=1}^T$ is normally chosen adaptively, by ensuring that the χ^2 divergence between successive distributions is constant and sufficiently small. The χ^2 divergence is approximated as $\chi^2(\mu_{n-1}|\mu_n) \approx \frac{N}{\mathrm{ESS}_n(\lambda_n)} - 1$ (see, e.g. Chopin and Papaspiliopoulos (2020, Section 8.6) for a justification), where ESS_n denotes the effective sample size

$$\mathrm{ESS}_n(\lambda) := \left(\sum_{i=1}^N w_n(\widetilde{X}_{n-1}^i)\right)^2 / \sum_{i=1}^N (w_n(\widetilde{X}_{n-1}^i))^2.$$

Given $\beta > 0$, in standard adaptive SMC we need to solve $\mathrm{ESS}_n(\lambda) = N/(\beta + 1)$ at each iteration n. This is normally achieved via the bisection method, since $\mathrm{ESS}_n(\lambda)$ is a nonlinear function of λ taking values in $[1, N]$.

We conducted a numerical experiment to study the possible behaviors of the tempering sequences found by such adaptive strategies when using SMC samplers. Our numerical results are obtained using waste-free SMC (Dau and Chopin, 2022), as there is evidence that it improves on standard SMC, which in turns outperforms annealed importance sampling (Jasra et al., 2011).

We use throughout the adaptive tempering SMC sampler as implemented in the package particles, with all its settings set to defaults; see https://github.com/nchopin/particles: the Markov kernels are random-walk Metropolis kernels, , automatically calibrated on the current particle sample; the next tempering exponent is chosen so that $\mathrm{ESS}_n = N/2$, and $N = 10^4$ and $d = 25$. The code is available at https://github.com/FrancescaCrucinio/MirrorDescentTempering.

Figure 2 (left) plots the tempering sequence $(\lambda_n)_{n\geq 1}$ computed adaptively on a toy example where $\mu_0 = \mathcal{N}_d(0_d, \mathrm{Id})$, $\pi = \mathcal{N}_d(m, \Sigma)$, $m = 1_d$, and various choices for Σ: (a) $\Sigma = 10^{-2}\,\mathrm{Id}$; (b) $\Sigma = 10^2\,\mathrm{Id}$; and (c) $\Sigma = \mathrm{diag}(v)$, with the first $(d/2)$ elements of v equal to 10^{-2}, and the remaining elements equal to 10^2. Cases (a) and (b) illustrates our theoretical tempering rates of Section 3.2; when the target has smaller (resp. larger) variance along all directions, the tempering sequence behaves like a positive (resp. negative) exponential. Case (c) is particularly interesting as it shows that the tempering sequence may behave as a mix between the two cases; when the variance of the target is both larger in certain directions and smaller in other directions (relative to μ_0), then the tempering sequence must slow down both at the beginning and at the end. The bottom line of this experiment is that what constitutes a "good" tempering sequence varies strongly according to the pair (μ_0, π), and thus using an adaptive strategy is essential for good performance.

Figure 2 (right) plots the number of tempering steps obtained from the algorithm as a function of d, in the "smaller variance" scenario, $\Sigma = 10^{-2}\,\mathrm{Id}$. One recovers the $\mathcal{O}(d^{1/2})$ scaling derived in Section 3.1. The reader may refer to Appendix E for more details on the implementation.

Remark 2. *Proposition 2 provides alternative methods to approximately solve $\chi^2(\mu_{n-1}|\mu_n) = \beta$ in a equivalent way (up to higher order terms) from the current set of particles.*

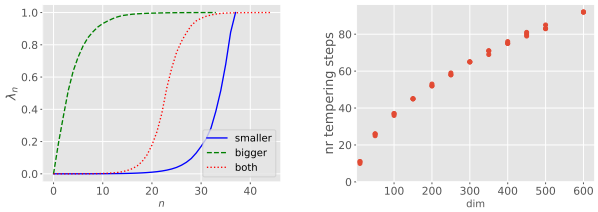

Figure 2: Left: adapted tempering sequences for different Σ. Right: Length of tempering sequence as a function of d in scenario (a); one recovers the $\mathcal{O}(d^{1/2})$ scaling.

A first one is to fix to a constant:

$$\mathrm{KL}(\mu_{n-1}|\mu_n) \approx -\frac{1}{N}\sum_{i=1}^{N} \log w_n(\widetilde{X}^i_{n-1}) + \log \frac{1}{N}\sum_{i=1}^{N} w_n(\widetilde{X}^i_{n-1}),$$

which is likely to be more stable than the ESS, since it involves the log-weights rather than the weights themselves. The second one is set the next λ_n as $\lambda_n = \lambda_{n-1} + (\beta/\hat{\mathrm{I}}(\lambda_{n-1}))^{1/2}$, where $\hat{\mathrm{I}}(\lambda_{n-1})$ is

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log \frac{\pi(\widetilde{X}^i_{n-1})}{\mu_0(\widetilde{X}^i_{n-1})}\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\log \frac{\pi(\widetilde{X}^i_{n-1})}{\mu_0(\widetilde{X}^i_{n-1})}\right)^2.$$

5. Related work

In this section we discuss alternative tempering strategies and algorithmic approximations related to the tempering update.

Tempering, KL divergence optimization and normalizing constant estimation.

The insight given by the mirror descent perspective allows us to relate sampling and estimation of the normalizing constant $\mathcal{Z}$ of π. In the AIS literature, the optimal sequence of distributions $(\mu_n)_{n\geq 1}$ is normally chosen to minimize the bias of the log-weights (Grosse et al., 2013; Goshtasbpour et al., 2023)

$$\log \mathcal{Z} - \mathbb{E}[\log w_n] = \sum_{n=1}^{T} \mathrm{KL}(\mu_{n-1}|\mu_n).$$

Assuming the first $(n-1)$ iterates are fixed, one finds μ_n by minimizing $\mathrm{KL}(\mu_{n-1}|\mu_n)$, which corresponds to the approach adopted in the SMC literature described above.

Grosse et al. (2013) derive optimal paths for exponential families and show that

$$\mu_\lambda = \arg\min_{\mu} \left[(1-\lambda)\,\mathrm{KL}(\mu|\mu_0) + \lambda\,\mathrm{KL}(\mu|\pi)\right].$$

This corresponds to the first step of entropic mirror descent, i.e. (3) with $n=0$ and $\lambda = \gamma_1$.

Goshtasbpour et al. (2023, Proposition 3.2) show that the tempering sequence is the path of steepest descent for the KL; i.e. that minimizes (1) infinitesimally, where the perturbation ξ is a smooth (C^1) perturbation with bounded variance. Given $(n-1)$ tempering iterates, they select the next one minimizing $\mathrm{KL}(\mu_n|\pi)$ instead and identify a tempering schedule that decreases this objective with constant rate and satisfies the ODE

$$\dot{\lambda} = c\,[\mathrm{I}(\lambda)(1-\lambda)]^{-1}. \tag{20}$$

This differs from ours in (16) which keeps the KL between successive entropic mirror descent iterates constant. Both strategies can be justified using the well known identity (Brekelmans et al. (2020, Section 4.4) and Appendix F)

$$\mathrm{KL}(\mu_\lambda|\mu_{\lambda'}) = \int_{\lambda}^{\lambda'} (\lambda - u)\mathrm{I}(u)du.$$

Using (14), we find that

$$0 \leq \mathrm{KL}(\mu_{n-1}|\mu_n) \leq \mathrm{KL}(\mu_{n-1}|\pi) - \mathrm{KL}(\mu_n|\pi),$$

which shows that the (16) and (20) fulfil opposite goals, i.e. (20) aims to find μ_n which minimizes $\mathrm{KL}(\mu_n|\pi)$ i.e. an upper bound on $\mathrm{KL}(\mu_{n-1}|\mu_n)$. From the numerical point of view, both strategies employ the importance weights to select the next λ; however, in our strategy the weights are those obtained by importance sampling with target μ_{λ_n} and proposal $\mu_{\lambda_{n-1}}$, in Goshtasbpour et al. (2023) the target is π. As a consequence, their method is more numerically unstable due to the higher variance of the weights. To see this we reproduce their narrow Gaussian experiment in Figure 3 and compare with waste-free SMC with the same setup of Section 4; we also include the results of SMC and AIS with the same setup but $\gamma = 0.05$ constant (see Appendix F.1 for implementation details). The target is $\pi = \mathcal{N}(1_d, 0.1^2\,\mathrm{Id})$ and $\mu_0 = \mathcal{N}(0_d, \mathrm{Id})$. Adaptive SMC better approximates π and requires only 5 tempering steps, while Goshtasbpour et al. (2023) provides worse approximations and require more than 6000 steps. Similarly, the algorithms with constant γ require more steps and provide worse approximations than adaptive SMC.

Kiwaki (2015) consider the variance of $\log w_n$ instead and derive an ODE similar to (16). Their method however requires running AIS twice to select the successive λ, while our rule (15) only requires to evaluate $\hat{\mathrm{I}}(\lambda)$

An ODE involving $\mathrm{I}(\lambda)$ was also derived in Gelman and Meng (1998), but, as mentioned by the authors, it often results in intractable optimal paths.

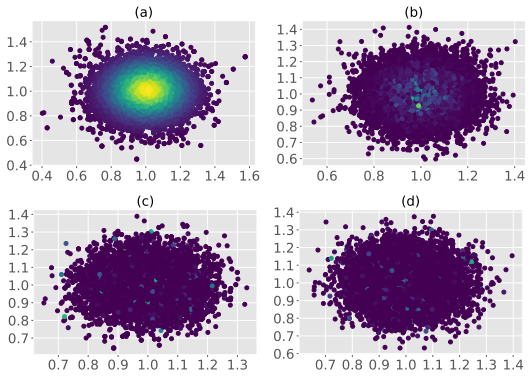

Figure 3: Approximations of $\pi = \mathcal{N}(1_d, 0.1^2\mathsf{Id})$. (a): adaptive SMC, 5 tempering steps, ESS = 0.79. (b): Goshtasbpour et al. (2023), 6257 tempering steps, ESS = 0.38. (c): SMC with constant step size $\gamma = 0.05$, 730 tempering steps, ESS = 0.99. (d): AIS with constant step size $\gamma = 0.05$, 730 tempering steps, ESS = 0.04.

Effect of the dimension on tempering/SMC in Beskos et al. (2014). In this paper the authors investigate the effect of dimension on the stability of SMC methods. There, stability refers to the ability of the SMC algorithm to produce accurate approximations of the target distribution as the dimension increases, while keeping the computational cost reasonable. The authors show that for a certain class of target densities (an i.i.d. target of the form $\pi(x) = \prod_{i=1}^d \pi^i(x^i)$ where $x = (x^1, \dots, x^d) \in \mathbb{R}^d$), SMC with the tempering sequence defined as $\lambda_n = \lambda_1 + (n-1)(1-\lambda_1)d^{-1}$, $1 \leq n \leq d$, is stable, i.e. the ESS converges weakly to a non-trivial limit (ESS $\in (1, N)$) as d grows and the number of particles is kept fixed. This result suggests using $\mathcal{O}(d)$ tempering iterations contrary to the $\mathcal{O}(d^{1/2})$ found in Proposition 2 and confirmed by our numerical results. We leave further investigation of the optimal scaling with d of SMC samplers for future work.

Parallel tempering. The tempering iterates (10) are also at the basis of Parallel Tempering (PT) (Geyer, 1991; Hukushima and Nemoto, 1996), a class of Markov chain Monte Carlo algorithms which relies on interacting Markov chains to sample from (10). PT is not based on importance sampling, hence the connection with Mirror Descent is less clear since identifying an update of the form (17) is not possible. It is customary in PT (Syed et al., 2022, e.g. Section 4) to fix the tempering sequences so that the acceptance probability of a swap between two successive λ_n is constant in n. One may use Proposition 1 of Predescu et al. (2004), see also Theorem 2 in Syed et al. (2022), which is similar in spirit to our Proposition 2, but not equivalent: Proposition 2 applies to the f divergence between μ_λ and $\mu_{\lambda'}$, for $\lambda' \approx \lambda$, where f is differentiable, whereas the acceptance rate of a PT swap is the TV distance between $\mu_\lambda \otimes \mu_{\lambda'}$ and $\mu_{\lambda'} \otimes \mu_\lambda$, again for $\lambda' \approx \lambda$. Moreover, the TV distance is a f-divergence with $f(t) = |t-1|$, which is not differentiable at 1.

Adaptive tempering. In Korba and Portier (2022), the authors propose to choose the step-size γ_n as follows. At time n draw m_n particles from q_n^{SRAIS}. Let $P = \sum_{l=1}^{m_n} u_n^l \delta_{X_n^l}$ and $Q = \sum_{l=1}^{m_n} (m_n)^{-1} \delta_{X_n^l}$ the reweighted and uniform distribution on the particles $(X_n^l)_{l=1}^{m_n}$ respectively, where $u_n^l = u_n(X_n^l) = \pi(x)/q_n(x)$ are the importance weights between the target distribution π and the current approximate iterate q_n of μ_n. Korba and Portier (2022) propose to set γ_n as $\gamma_n = 1 - R_\alpha(P|Q)/\log(m_n)$, where R_α is Renyi's α-divergence (Rényi et al., 1961) of P from Q, in particular $R_1(P||Q) = \mathrm{KL}(P|Q)$, and $\log(m_n)$ normalizes the ratio between 0 and 1. Hence, for $\alpha = 1$ and without the discrete particle approximation, their rule can be written $\gamma_n = 1 - \mathrm{KL}(\mu_n|\pi)$. Since $\gamma_n = (\lambda_n - \lambda_{n-1})/(1-\lambda_{n-1})$, by the decrease of KL formula (14), this rule can also be seen as enforcing a gap between consecutive λ's as a constant.

6. Conclusion

This paper establishes a connection between entropic mirror descent and tempering to sample from a target probability distribution known up to a normalizing constant. We show that the two strategies are equivalent and obtain an explicit convergence rate for the tempering iterates. This convergence rate does not depend on the convexity properties of the target π, contrary to the rates for Langevin Monte Carlo (Durmus et al., 2019; Vempala and Wibisono, 2019; Karimi et al., 2016).

We provide an optimization point of view on sequential Monte Carlo by identifying the SMC update as (17) and motivate the adaptive strategy commonly used in the literature through mirror descent. Furthermore, we identify that for a number of algorithms based on importance sampling, the importance weights carry the gradient information, and propose strategies to reduce their numerical error (see Appendix E). By comparing several approximations of entropic mirror descent and several adaptive strategies to select the sequence, we find that SMC has generally lower cost and the tempering rule (16) is more stable than alternatives. This connection enabled us to tackle the selection of the tempering schedule in a principled way, and derive several conclusions that were not known in or contradicts the previous tempering literature, for instance that the length of the tempering schedule should scale as $\sqrt{d}$. Our approach yields a simpler and more numerically stable tempering rule than other schemes (minus the standard ESS−based rule in the SMC literature, which gives essentially the same results as ours).

Impact statement

Our paper has a theoretical and practical interest for the litterature on sampling and MCMC algorithms. On the theoretical side, our study yields a rate of convergence for the target sequence that SMC algorithms are tracking. On the practical side, we show that common practices such as the ESS-based rules, are more theoretically grounded, simpler and more efficient than alternative proposals. This may have a substantial impact in the deployment of Bayesian inference tasks which rely on MCMC algorithms, and enable to predict with uncertainty.

References

Amari, S.-i. (2016). *Information Geometry and Its Applications*, volume 194. Springer.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21:555–568.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. (2022). Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Beskos, A., Crisan, D., and Jasra, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Annals of Applied Probability*, 24(4):1396–1445.

Brekelmans, R., Masrani, V., Wood, F., Steeg, G. V., and Galstyan, A. (2020). All in the exponential family: Bregman duality in thermodynamic variational inference. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1111–1122. PMLR.

Chacón, J. E. and Duong, T. (2018). *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC.

Chen, Y., Huang, D. Z., Huang, J., Reich, S., and Stuart, A. M. (2023). Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*.

Chizat, L. (2022). Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures. *Open Journal of Mathematical Optimization*, 3.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer.

Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746.

Dai, B., He, N., Dai, H., and Song, L. (2016). Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR.

Dai, C., Heng, J., Jacob, P. E., and Whiteley, N. (2022). An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association*, 117(539):1587–1600.

Dau, H.-D. and Chopin, N. (2022). Waste-free sequential Monte Carlo. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):114–148.

Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer Verlag, New York.

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

Domingo-Enrich, C. and Pooladian, A.-A. (2023). An Explicit Expansion of the Kullback-Leibler Divergence along its Fisher-Rao Gradient Flow. *Transactions on Machine Learning Research*.

Durmus, A., Majewski, S., and Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711.

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

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185.

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

Goshtasbpour, S., Cohen, V., and Perez-Cruz, F. (2023). Adaptive annealed importance sampling with constant rate progress. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings*

of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 11642–11658. PMLR.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. (2013). Annealing between distributions by averaging moments. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Hukushima, K. and Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608.

Iba, Y. (2001). Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, 12(05):623–656.

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22.

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

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer.

Karimi, M. R., Hsieh, Y.-P., and Krause, A. (2023). Sinkhorn flow: A continuous-time framework for understanding and generalizing the sinkhorn algorithm. *arXiv preprint arXiv:2311.16706*.

Kiwaki, T. (2015). Variational optimization of annealing schedules. *arXiv preprint arXiv:1502.05313*.

Korba, A. and Portier, F. (2022). Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354.

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

Mou, W., Ma, Y.-A., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. (2021). High-order Langevin diffusion yields an accelerated MCMC algorithm. *The Journal of Machine Learning Research*, 22(1):1919–1959.

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

Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400.

Predescu, C., Predescu, M., and Ciobanu, C. V. (2004). The incomplete beta function law for parallel tempering sampling of classical canonical systems. *The Journal of Chemical Physics*, 120(9):4119–4128.

Rényi, A. et al. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Salim, A., Korba, A., and Luise, G. (2020). The Wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2022). Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350.

Syed, S., Romaniello, V., Campbell, T., and Bouchard-Côté, A. (2021). Parallel tempering on optimized paths. In *International Conference on Machine Learning*, pages 10033–10042. PMLR.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32.

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

Yan, Y., Wang, K., and Rigollet, P. (2023). Learning Gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow. *arXiv preprint arXiv:2301.01766*.

Ying, L. (2020). Mirror descent algorithms for minimizing interacting free energy. *Journal of Scientific Computing*, 84(3):51.

A. Proof of Proposition 1

The proof below is written for a generic L-smooth and l-strongly convex functional $\mathcal{F}$ relatively to a Bregman potential ϕ. Recall that in the case of $\mathcal{F}$ being the Kullback-Leibler divergence and negative entropy, $L = l = 1$.

We first state a preliminary result, known as the "three-point inequality" or "Bregman proximal inequality", which can be found in Aubin-Frankowski et al. (2022, Lemma 3).

Lemma 1 (Three-point inequality)**.** *Given $\mu \in \mathcal{M}(\mathbb{R}^d)$ and some proper convex functional $\mathcal{G} : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$, if $\nabla\phi(\mu)$ exists, as well as $\bar{\nu} = \operatorname{argmin}_{\nu \in C}\{\mathcal{G}(\nu) + B_\phi(\nu|\mu)\}$, then for all $\nu \in C \cap \operatorname{dom}(\phi) \cap \operatorname{dom}(\mathcal{G})$:*

$$\mathcal{G}(\nu) + B_\phi(\nu|\mu) \geq \mathcal{G}(\bar{\nu}) + B_\phi(\bar{\nu}|\mu) + B_\phi(\nu|\bar{\nu}). \tag{21}$$

We can now start the proof of mirror descent convergence rate. Since $\mathcal{F}$ is L-smooth relative to ϕ and $\gamma_{n+1} < 1/L$ implies, we have

$$\begin{aligned}\mathcal{F}(\mu_{n+1}) &\leq \mathcal{F}(\mu_n) + \langle \nabla\mathcal{F}(\mu_n), \mu_{n+1} - \mu_n\rangle + LB_\phi(\mu_{n+1}|\mu_n) \\ &\leq \mathcal{F}(\mu_n) + \langle \nabla\mathcal{F}(\mu_n), \mu_{n+1} - \mu_n\rangle + \frac{1}{\gamma_{n+1}} B_\phi(\mu_{n+1}|\mu_n).\end{aligned} \tag{22}$$

Applying Lemma 1 to the convex function $\mathcal{G}_n(\nu) = \gamma_{n+1}\langle \nabla\mathcal{F}(\mu_n), \nu - \mu_n\rangle$, with $\mu = \mu_n$ and $\bar{\nu} = \mu_{n+1}$ yields

$$\langle \nabla\mathcal{F}(\mu_n), \mu_{n+1} - \mu_n\rangle + \frac{1}{\gamma_{n+1}} B_\phi(\mu_{n+1}|\mu_n) \leq \langle \nabla\mathcal{F}(\mu_n), \nu - \mu_n\rangle + \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_{n+1}).$$

Fix ν, then (22) becomes:

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) + \langle \nabla\mathcal{F}(\mu_n), \nu - \mu_n\rangle + \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_{n+1}). \tag{23}$$

This shows in particular, by substituting $\nu = \mu_n$ and since $B_\phi(\nu|\mu_{n+1}) \geq 0$, that

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) - \frac{1}{\gamma_{n+1}} B_\phi(\mu_n|\mu_{n+1}), \tag{24}$$

i.e. $\mathcal{F}$ is decreasing at each iteration. Since $\mathcal{F}$ is l-strongly convex relative to ϕ, we also have:

$$\langle \nabla\mathcal{F}(\mu_n), \nu - \mu_n\rangle \leq \mathcal{F}(\nu) - \mathcal{F}(\mu_n) - lB_\phi(\nu|\mu_n)$$

and (23) becomes:

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\nu) + \left(\frac{1}{\gamma_{n+1}} - l\right) B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_{n+1}), \tag{25}$$

i.e., multiplying the previous equation by $(\gamma_{n+1}^{-1} - l)^{-1}$, we get

$$\left(\frac{1}{1 - \gamma_{n+1}l}\right) \mathcal{F}(\mu_{n+1}) \leq \left(\frac{1}{1 - \gamma_{n+1}l}\right) \mathcal{F}(\nu) + \frac{1}{\gamma_{n+1}} B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}} \left(\frac{1}{1 - \gamma_{n+1}l}\right) B_\phi(\nu|\mu_{n+1}). \tag{26}$$

A.1. Proof for $(\gamma_n)_{n\geq 1}$ decreasing or constant

Similarly to Lu et al. (2018), we now consider for $n \geq 1$:

$$\mathcal{P}(n): \quad \sum_{k=1}^{n} \left(\frac{1}{1 - \gamma_k l}\right)^k \mathcal{F}(\mu_k) \leq \sum_{k=1}^{n} \left(\frac{1}{1 - \gamma_k l}\right)^k \mathcal{F}(\nu) + \frac{1}{\gamma_1} B_\phi(\nu|\mu_0) - \frac{1}{\gamma_n} \left(\frac{1}{1 - \gamma_n l}\right)^n B_\phi(\nu|\mu_n).$$

We first have that

$$\mathcal{P}(1): \left(\frac{1}{1 - \gamma_1 l}\right) \mathcal{F}(\mu_1) \leq \left(\frac{1}{1 - \gamma_1 l}\right) \mathcal{F}(\nu) + \frac{1}{\gamma_1} B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1} \left(\frac{1}{1 - \gamma_1 l}\right) B_\phi(\nu|\mu_1)$$

is true by (26). Then, assume $\mathcal{P}(n)$ holds. We have by (26):

$$\begin{aligned}\sum_{k=1}^{n+1}\left(\frac{1}{1-\gamma_k l}\right)^k \mathcal{F}(\mu_k) &= \sum_{k=1}^{n}\left(\frac{1}{1-\gamma_k l}\right)^k \mathcal{F}(\mu_k) + \left(\frac{1}{1-\gamma_{n+1} l}\right)^{n+1} \mathcal{F}(\mu_{n+1}) \\ &\leq \sum_{k=1}^{n}\left(\frac{1}{1-\gamma_k l}\right)^k \mathcal{F}(\nu) + \frac{1}{\gamma_1} B_\phi(\nu|\mu_0) - \frac{1}{\gamma_n}\left(\frac{1}{1-\gamma_n l}\right)^n B_\phi(\nu|\mu_n) \\ &+ \left(\frac{1}{1-\gamma_{n+1} l}\right)^{n+1} \mathcal{F}(\nu) + \frac{1}{\gamma_{n+1}}\left(\frac{1}{1-\gamma_{n+1} l}\right)^n B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}}\left(\frac{1}{1-\gamma_{n+1} l}\right)^{n+1} B_\phi(\nu|\mu_{n+1}) \\ &\leq \sum_{k=1}^{n+1}\left(\frac{1}{1-\gamma_k l}\right)^k \mathcal{F}(\nu) + \frac{1}{\gamma_1} B_\phi(\nu|\mu_0) - \frac{1}{\gamma_{n+1}}\left(\frac{1}{1-\gamma_{n+1} l}\right)^{n+1} B_\phi(\nu|\mu_{n+1})\end{aligned}$$

where we used in the last inequality (to upper bound the sum of the third and fifth term by zero) that $s \mapsto s^{-1}(1-s)^{-n}$ was a monotone increasing function and that the sequence $(\gamma_n)_{n\geq 1}$ was decreasing, showing $\mathcal{P}(n+1)$ holds. Hence $\mathcal{P}(n)$ is true for all $n \geq 1$. Then, using the monotonicity of $(\mathcal{F}(\mu_n))_{n\geq 0}$ on the left hand side and the positivity of $B_\phi(\nu|\mu_n)$ on the right hand side of $\mathcal{P}(n)$, we have

$$\sum_{k=1}^{n}\left(\frac{1}{1-\gamma_k l}\right)^k (\mathcal{F}(\mu_n) - \mathcal{F}(\nu)) \leq \frac{1}{\gamma_1} B_\phi(\nu|\mu_0) - \frac{1}{\gamma_n}\left(\frac{1}{1-\gamma_n l}\right)^n B_\phi(\nu|\mu_n) \leq \frac{1}{\gamma_1} B_\phi(\nu|\mu_0).$$

This shows that

$$\mathcal{F}(\mu_n) - \mathcal{F}(\mu) \leq \frac{C_n}{\gamma_1} B_\phi(\nu|\mu_0), \text{ where } C_n^{-1} = \sum_{k=1}^{n}\left(\frac{1}{1-\gamma_k l}\right)^k.$$

A.2. Proof for general $(\gamma_n)_{n\geq 1}$

Consider for $n \geq 1$:

$$\mathcal{P}(n): \quad \sum_{k=1}^{n}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\mu_k) \leq \sum_{k=1}^{n}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\nu) + \frac{1}{\gamma_1}B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1}\prod_{i=1}^{n}\frac{1}{1-\gamma_i l}B_\phi(\nu|\mu_n)$$

We first have that

$$\mathcal{P}(1): \left(\frac{1}{1-\gamma_1 l}\right)\mathcal{F}(\mu_1) \leq \left(\frac{1}{1-\gamma_1 l}\right)\mathcal{F}(\nu) + \frac{1}{\gamma_1}B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1}\left(\frac{1}{1-\gamma_1 l}\right)B_\phi(\nu|\mu_1)$$

is true by (26). Then, assume $\mathcal{P}(n)$ holds. We have by (26):

$$\begin{aligned}\sum_{k=1}^{n+1}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\mu_k) &= \sum_{k=1}^{n}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\mu_k) + \frac{\gamma_{n+1}}{\gamma_1}\prod_{i=1}^{n+1}\frac{1}{1-\gamma_i l}\mathcal{F}(\mu_{n+1}) \\ &\leq \sum_{k=1}^{n}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\nu) + \frac{1}{\gamma_1}B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1}\prod_{i=1}^{n}\frac{1}{1-\gamma_i l}B_\phi(\nu|\mu_n) \\ &+ \frac{\gamma_{n+1}}{\gamma_1}\prod_{i=1}^{n}\frac{1}{1-\gamma_i l}\left(\left(\frac{1}{1-\gamma_{n+1} l}\right)\mathcal{F}(\nu) + \frac{1}{\gamma_{n+1}}B_\phi(\nu|\mu_n) - \frac{1}{\gamma_{n+1}}\left(\frac{1}{1-\gamma_{n+1} l}\right)B_\phi(\nu|\mu_{n+1})\right) \\ &= \sum_{k=1}^{n+1}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}\mathcal{F}(\nu) + \frac{1}{\gamma_1}B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1}\prod_{i=1}^{n+1}\frac{1}{1-\gamma_i l}B_\phi(\nu|\mu_{n+1}),\end{aligned}$$

showing $\mathcal{P}(n+1)$ holds. Hence $\mathcal{P}(n)$ is true for all $n \geq 1$. Then, using the monotonicity of $(\mathcal{F}(\mu_n))_{n\geq 0}$ on the left hand side and the positivity of $B_\phi(\nu|\mu_n)$ on the right hand side of $\mathcal{P}(n)$, we have

$$\sum_{k=1}^{n}\frac{\gamma_k}{\gamma_1}\prod_{i=1}^{k}\frac{1}{1-\gamma_i l}(\mathcal{F}(\mu_n) - \mathcal{F}(\nu)) \leq \frac{1}{\gamma_1}B_\phi(\nu|\mu_0) - \frac{1}{\gamma_1}\prod_{i=1}^{n}\frac{1}{1-\gamma_i l}B_\phi(\nu|\mu_n) \leq \frac{1}{\gamma_1}B_\phi(\nu|\mu_0).$$

This shows that

$$\mathcal{F}(\mu_n) - \mathcal{F}(\mu) \leq \frac{C_n}{\gamma_1} B_\phi(\nu|\mu_0), \text{ where } C_n^{-1} = \sum_{k=1}^n \frac{\gamma_k}{\gamma_1} \prod_{i=1}^k \frac{1}{1-\gamma_i l}.$$

In particular for $l = 1$ and $\gamma_k = (\lambda_k - \lambda_{k-1})/(1-\lambda_{k-1})$,

$$C_n^{-1} = \frac{1}{\lambda_1} \sum_{k=1}^n \left(\frac{\lambda_k - \lambda_{k-1}}{1-\lambda_{k-1}} \frac{1-\lambda_0}{1-\lambda_k} \right) = \frac{1}{\lambda_1} \sum_{k=1}^n \left(\frac{\lambda_k - \lambda_{k-1}}{(1-\lambda_{k-1})(1-\lambda_k)} \right),$$

where we used $\lambda_0 = 0$.

A.3. Bounds on convergence rate

In this section we prove upper bounds on the convergence rates previously obtained. Our bounds are obtained in the case $l = 1$ and $\gamma_n \leq 1/L = 1$ for all $n \geq 1$.

We show that

$$C_n = \left(\sum_{k=1}^n \frac{\gamma_k}{\gamma_1} \prod_{i=1}^k \frac{1}{1-\gamma_i} \right)^{-1} \leq \prod_{k=1}^n (1-\gamma_k). \tag{27}$$

To see this we consider for $n \geq 1$, $\mathcal{P}(n)$: $\sum_{k=1}^n \frac{\gamma_k}{\gamma_1} \prod_{i=1}^k \frac{1}{1-\gamma_i} \geq \prod_{k=1}^n (1-\gamma_k)^{-1}$. We trivially have that $\mathcal{P}(1)$: $\left(\frac{1}{1-\gamma_1}\right)^1 \geq (1-\gamma_1)^{-1}$ is true. Then, assume $\mathcal{P}(n)$ holds. We have

$$\begin{aligned}\sum_{k=1}^{n+1} \frac{\gamma_k}{\gamma_1} \prod_{i=1}^k \frac{1}{1-\gamma_i} &= \sum_{k=1}^n \frac{\gamma_k}{\gamma_1} \prod_{i=1}^k \frac{1}{1-\gamma_i} + \frac{\gamma_{n+1}}{\gamma_1} \prod_{i=1}^{n+1} \frac{1}{1-\gamma_i} \geq \prod_{k=1}^n (1-\gamma_k)^{-1} + \gamma_{n+1} \prod_{k=1}^{n+1} (1-\gamma_k)^{-1} \\ &= \prod_{k=1}^n (1-\gamma_k)^{-1} \left[1 + \gamma_{n+1}(1-\gamma_{n+1})^{-1}\right] = \prod_{k=1}^{n+1} (1-\gamma_k)^{-1}\end{aligned}$$

since $\gamma_1 \leq 1$ for all $n \geq 1$, showing $\mathcal{P}(n+1)$ holds. Hence (27) is true for all $n \geq 1$.

B. Proof of Proposition 2

B.1. Tempering sequence as a parametric model

Let us recall that the tempering sequence is defined as:

$$\mu_\lambda(x) = \frac{\mu_0^{1-\lambda}(x)\pi^\lambda(x)}{\exp\{\psi(\lambda)\}} = \mu_0(x) \exp\{\lambda s(x) - \psi(\lambda)\}$$

for $\lambda \in [0,1]$, where $s(x) := \log \pi(x)/\mu_0(x)$, and

$$\psi(\lambda) := \log \int \mu_0(x) \exp\{\lambda s(x)\}\, dx$$

is the partition function (log-normalizing constant).

In our case, the Fisher's score is: $t_\lambda(x) = s(x) - \psi'(\lambda)$ (the score has expectation zero, as expected since $\psi'(\lambda) = \mathbb{E}_\lambda[s(X)]$), and

$$I(\lambda) := \mathrm{Var}_\lambda\left[t_\lambda(X)\right] = \mathbb{E}_\lambda\left[t_\lambda(X)^2\right] = \mathrm{Var}_\lambda\left[s(X)\right].$$

Note also the well-known identity:

$$I(\lambda) = -\mathbb{E}_\lambda[t'_\lambda(X)] = -\mathbb{E}_\lambda\left[\frac{\partial^2 \log \mu_\lambda(X)}{\partial \lambda^2}\right].$$

B.2. Proof

Recall that f must be convex and such that $f(1) = 0$.

By the standard properties of $f-$divergence, it is clear that the function $\varphi_\lambda : \lambda' \to D_f(\lambda'|\lambda)$ is non-negative, and zero at $\lambda' = \lambda$, hence its first derivative must be zero at $\lambda' = \lambda$. In fact,

$$\varphi'_\lambda(\lambda') = \int \mu'_{\lambda'} f'(\mu_{\lambda'}/\mu_\lambda) = \int \mu_{\lambda'} t_{\lambda'} f'(\mu_{\lambda'}/\mu_\lambda)$$

and note we have indeed $\varphi'_\lambda(\lambda) = f'(1) \int \mu'_\lambda = 0$. For the second derivative

$$\varphi''_\lambda(\lambda') = \int \mu_{\lambda'}(t'_{\lambda'} + t^2_{\lambda'}) f'\left(\frac{\mu_{\lambda'}}{\mu_\lambda}\right) + \int \frac{(\mu_{\lambda'} t_{\lambda'})^2}{\mu_\lambda} f''\left(\frac{\mu_{\lambda'}}{\mu_\lambda}\right)$$

and at $\lambda' = \lambda$:

$$\begin{aligned}\varphi''_\lambda(\lambda) &= f'(1) \int \mu_\lambda(t'_\lambda + t^2_\lambda) + f''(1) \int \mu_\lambda \left(t_\lambda\right)^2 \\ &= f''(1) I(\lambda).\end{aligned}$$

This ends the proof.

C. Derivation of mirror descent

To obtain (4) write the first order conditions of (3), i.e. let $\mathcal{G}(\mu) = \langle \nabla \mathcal{F}(\mu_n), \mu - \mu_n \rangle + B_\phi(\mu|\mu_n)$ where $B_\phi(\mu|\mu_n)$ is given in (2), differentiate w.r.t. μ and set the differential of $\mathcal{G}$ to 0.

To obtain the entropic mirror descent update (6), notice that if in (4), ϕ is chosen as the negative entropy (5) where $\nabla\phi(\mu) = \log(\mu)$. Then (4) becomes $\log(\mu_{n+1}) = \log(\mu_n) - \gamma_{n+1}\nabla\mathcal{F}(\mu_n)$. Exponentiating both sides we obtain (6) since μ_{n+1} is constrained to be a probability distribution; as we only know that the density is proportional to the r.h.s. of (6), it should be renormalized accordingly.

When choosing the objective as a KL, we relate this scheme to tempering sequences through (7). However, we note that other choices of KL objectives enable to relate this scheme to other algorithms, such as Expectation-Maximization or Sinkhorn's algorithm (Aubin-Frankowski et al., 2022; Karimi et al., 2023).

D. Entropic mirror descent is a time-discretization of Fisher-Rao flow

Entropic mirror descent iteration on $\mathcal{F}$ starting from μ_0, is an Euler (or Forward) time-discretisation of the Fisher-Rao flow of $\mathcal{F}$ (Lu et al., 2023). Indeed, the FR flow of a functional $\mathcal{F}$ can be written

$$\frac{\partial \mu_t}{\partial t} = -\mu_t \mathcal{F}'(\mu_t), \text{ hence, } \frac{\partial \log(\mu_t)}{\partial t} = -\mathcal{F}'(\mu_t).$$

An Euler discretization of the previous continuous dynamics write:

$$\log(\mu_{l+1}) - \log(\mu_l) = -\gamma_{l+1}\mathcal{F}'(\mu_l) \ , \tag{28}$$

which recovers (6) by exponentiating the equality.

E. Algorithms details

We collect here further details on the SMC samplers described in Section 4 and describe other strategies based on importance sampling that approximate the mirror descent iterates (6).

E.1. Other schemes

Particle Mirror Descent (PMD). Similarly to SMC, Dai et al. (2016) propose an approximation of the mirror descent iterates (7) based on importance sampling. The mirror descent iterate at time n is approximated by a kernel density estimator

Algorithm 1 SMC samplers (Del Moral et al., 2006).

Inputs: sequences of temperatures $(\lambda_n)_{n=1}^T$, Markov kernels $(M_n)_{n=1}^T$, initial proposal μ_0.
Initialize: set $\lambda_0 = 0$, 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
  Propose: draw $X_n^i \sim M_n(\cdot, \widetilde{X}_{n-1}^i)$ for $i = 1, \dots, N$.
  Reweight: compute and normalize the weights $W_n^i \propto w_n(\widetilde{X}_{n-1}^i)$ for $i = 1, \dots, N$.
end for
Output: $q_n(x) = \sum_{i=1}^N W_n^i \delta_{X_n^i}(x)$

(KDE)

$$q_n^{\text{PMD}}(x) := \sum_{i=1}^N V_n^i K_{h_n}(x - X_n^i), \tag{29}$$

where $\{X_n^i, V_n^i\}_{i=1}^N$ denotes a weighted particle set and K_{h_n} is a smoothing kernel with bandwidth h_n. At iteration n the weighted particle set $\{X_{n-1}^i, V_{n-1}^i\}_{i=1}^N$ is resampled to obtain the equally weighted particle set $\{\widetilde{X}_{n-1}^i, 1/N\}_{i=1}^N$, the kernel K_{h_n} is applied to propose new particle locations $X_n^i \sim K_{h_n}(\cdot - \widetilde{X}_{n-1}^i)$. The weights for the proposed particle set are then proportional to

$$v_n(x) = \left(\frac{\pi(x)}{q_{n-1}^{\text{PMD}}(x)}\right)^{\gamma_n}. \tag{30}$$

It is then clear that the Particle Mirror Descent (PMD) scheme summarized in Algorithm 2 in the Appendix is of the form (17).

Comparing PMD with SMCand with the vast literature on SMC algorithms (see, e.g., Chopin and Papaspiliopoulos (2020) for a comprehensive introduction) we find that PMD is an SMC algorithm targeting the sequence of distributions

$$\tilde{\mu}_n(x) \propto \int \mu_{n-1}(x') K_{h_n}(x - x') dx' \left(\frac{\pi(x)}{\eta_{n-1}(x)}\right)^{\gamma_n}, \tag{31}$$

which converges to the mirror descent iterates (7) as $h_n \to 0$. The kernel K_{h_n} is replacing the μ_n-invariant kernel M_n as proposal and the importance weights are given by (30). However, PMD is not a standard SMC algorithm, since the weights v_n are approximations of the idealized weights $v_n^\star(x) = \pi(x) / \int K_{h_n}(x - x')\mu_{n-1}(x')dx'$ obtained by plugging the KDE q_{n-1}^{PMD} in place of the denominator. Hence PMD uses one more approximation than standard SMC samplers.

Leveraging the connection between mirror descent and tempering established in Section 3, it is easy to see that $v_n^\star \to w_n$ as $h_n \to 0$ (see (19)). Hence, we could replace v_n with w_n in Algorithm 2 to reduce its computational cost and numerical error, since v_n requires an $\mathcal{O}(N)$ cost due to the presence of the kernel density estimator q_{n-1}, while the cost of w_n is $\mathcal{O}(1)$. Nevertheless, this does not lead to an SMC algorithm targeting $\tilde{\mu}_n$ (or μ_n).

Safe and Regularized Adaptive Importance Sampling (SRAIS). Korba and Portier (2022)) propose an algorithm detailed in Algorithm 3 in the Appendix, that samples at each iteration a particle X_n from a proposal q_n^{SRAIS}. Similarly to PMD which relies on the KDE estimator (29), SRAIS relies on a KDE estimate to approximate the mirror descent iterates

$$q_n^{\text{SRAIS}}(x) = \sum_{i=1}^n U_i K_{h_i}(x - X_i), \tag{32}$$

where $\{X_i, U_i\}_{i=1}^n$ denotes a weighted particle set. However, in this case the size of the particle population is not fixed and the KDE estimate uses all particles from previous iterations. Notice that the particle sampling step (Step 4 of Algorithm 3) can be repeated, resulting in sampling a batch m_n of particles at step n. The weights for the proposed particles are

$$u_n(x) = \left(\frac{\pi(x)}{q_{n-1}^{\text{SRAIS}}(x)}\right)^{\gamma_n}, \tag{33}$$

Algorithm 2 Particle Mirror Descent (PMD; Dai et al. (2016)).

Inputs: sequences of bandwidths $(h_n)_{n=1,\dots,T}$, learning rates $(\gamma_n)_{n=1,\dots,T}$, 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 $\{X_{n-1}^i, V_{n-1}^i\}_{i=1}^N$ and set $V_n^i = 1/N$ for $i = 1, \dots, N$.
  end if
  Propose: draw $X_n^i \sim K_{h_n}(\cdot - \widetilde{X}_{n-1}^i)$ for $i = 1, \dots, N$.
  Reweight: compute and normalize the weights $V_n^i \propto v_n(X_n^i)$ for $i = 1, \dots, N$.
end for
Output: $q_n(x) = \sum_{i=1}^N V_n^i K_{h_n}(x - X_n^i)$.

and we can identify SRAIS to be of the form (17). Similarly to PMD, one could replace u_n with w_n in SRAIS.

Remark 3. *In the original scheme proposed by Korba and Portier (2022), the proposal at each iteration is defined as* $\tilde{q}_{n+1} = (1 - r_{n+1})q_{n+1} + r_{n+1}q_0$ *where* q_{n+1} *is the KDE (32),* $(r_n)_{n\geq 0}$ *is a sequence in* $[0, 1]$ *converging to* 0 *and* q_0 *is a "safe" density (e.g. with heavy tails) preventing the importance weights from degeneracy. In Algorithm 3 we removed the dependency with the safe density and took the sequence* $(r_n)_{n\geq 0}$ *constant equal to zero for a clearer presentation.*

Algorithm 3 Safe and Regularized Adaptive Importance sampling (SRAIS; Korba and Portier (2022))

Inputs: Sequences of bandwidths $(h_n)_{n=1,\dots,T}$, learning rates $(\gamma_n)_{n=1,\dots,T}$, initial proposal μ_0.
Initialize: sample $X_1 \sim \mu_0$ and set $U_1 = (\pi(X_1)/\mu_0(X_1))^{\gamma_1}$.
for $n = 1, \dots, T$ **do**
  Propose: draw $X_{n+1} \sim q_n$
  Reweight: compute the weight $U_{n+1} \propto u_{n+1}(X_{n+1})$ and normalize the weights.
  Update the proposal as in (32).
end for
Output: $q_{n+1}(x) = \sum_{i=1}^{n+1} U^i K_{h_i}(x - X_i)$.

E.2. Comparison of algorithms

As discussed in the previous sections, both SMC samplers and PMD are an instance of SMC algorithms (albeit not a standard one in the case of PMD). The convergence properties of SMC defined in Algorithm 1 are guaranteed by the wide literature on SMC algorithms (see, e.g., Del Moral (2004) for a complete account). In particular, one can show (see Del Moral (2004, Theorem 7.4.3) and Crisan and Doucet (2002)) that every measurable bounded function $\varphi : \mathbb{R}^d \to \mathbb{R}$ with $\|\varphi\| := \sup_{x\in\mathbb{R}^d} |\varphi(x)| < \infty$,

$$\mathbb{E}\left[\left|\int \varphi \mathrm{d}\mu_n - \int \varphi \mathrm{d}q_n^{\mathrm{SMC}}\right|\right] \leq \frac{B_n^{\mathrm{SMC}}\|\varphi\|}{N^{1/2}}$$

where B_n^{SMC} denotes a finite constant which does not depend on N.

A similar result for Algorithm 2 has been established in Dai et al. (2016, Theorem 5). The approximation error of PMD is divided into an optimisation error, due to the fact that the algorithm is stopped at time T, and the following approximation error arising from the particle approximation to the target $\tilde{\mu}_n$ in (31)

$$\mathbb{E}\left[\left|\int \varphi \mathrm{d}\tilde{\mu}_n - \int \varphi \mathrm{d}q_n^{\mathrm{PMD}}\right|\right] \leq \frac{B_n^{\mathrm{PMD}}\|\varphi\|}{N^{1/2}},$$

where B_n^{SMC} denotes a finite constant which does not depend on N.

In the case of SMC samplers, there is no optimisation error since Algorithm 1 targets μ_n directly (and not the smoothed version (31)) and, by construction, at time T we have $\lambda_T = 1$ so that $\mu_T = \pi$.

Furthermore, when implementing Algorithm 1 there is no need to introduce the kernel K_{h_n} to obtain a KDE at each iteration, this results in a simpler algorithm than Algorithm 2 which does not require the bandwidth parameter h_n whose tuning is

notoriously difficult (Silverman, 1986). Additionally, KDE performs poorly if the dimension of the underlying space is large (Chacón and Duong, 2018).

The presence of the KDE in PMD also causes the algorithm to have a higher computational cost than standard SMC samplers, in fact, the presence of the KDE in the weights (30) means that these weights require an $\mathcal{O}(N)$ cost to be computed for each particle, against the $\mathcal{O}(1)$ per particle of the weights (18). These results in a $\mathcal{O}(NT)$ cost for Algorithm 1 and $\mathcal{O}(N^2T)$ for Algorithm 2. Clearly, the $\mathcal{O}(N^2T)$ of PMD could be reduced to $\mathcal{O}(NT)$ by replacing the weights (30) with (18), since the former are an approximation of the idealized weights $v_n^\star(x) = (\pi(x)/\mu_{n-1}(x))^{\gamma_n}$ which are proportional to (18) as shown in (19), at the cost of targeting a slightly different distribution.

The computational cost of iteration n of SRAIS is $\mathcal{O}(n)$ because of the KDE in the weights (33). Hence, the cost of Algorithm 3 is $\sum_{n=1}^{T}\mathcal{O}(n) \approx \mathcal{O}(T^2)$. In practice, to reduce computational cost, one could use only the last iterations as the first ones can be considered as "burn-in" steps.

F. Further discussion on (20) and implementation details

Consider the well-known identity Brekelmans et al. (2020, Section 4.4)

$$\begin{aligned}\mathrm{KL}(\mu_{\lambda_{n-1}}|\mu_{\lambda_n}) &= \int_{\lambda_{n-1}}^{\lambda_n}(\lambda_n-\lambda)\mathrm{Var}_\lambda\left[s(X)\right]d\lambda \\ &= \int_{\lambda_{n-1}}^{\lambda_n}(\lambda_n-\lambda)\mathrm{I}(\lambda)d\lambda.\end{aligned} \tag{34}$$

We want to study the infinitesimal behaviour of the KL when $\lambda_{n-1} = \lambda(t)$ but λ_n is fixed. As suggested in Goshtasbpour et al. (2023) a natural requirement is to keep the derivative of the KL w.r.t. time constant

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\mu_{\lambda(t)}|\mu_{\lambda_n}) &= \frac{d}{dt}\left(\int_{\lambda(t)}^{\lambda_n}(\lambda_n-\lambda)\mathrm{I}(\lambda)d\lambda\right) \\ &= \frac{d\lambda(t)}{dt}(\lambda_n-\lambda(t))\mathrm{I}(\lambda(t)) = c,\end{aligned}$$

where we used Leibniz integral rule for differentiation under the integral sign under the assumptions that all quantities are well-defined. This gives us the following ODE for $\lambda(t)$

$$\frac{d\lambda(t)}{dt} = c\left[(\lambda_n-\lambda(t))\mathrm{I}(\lambda(t))\right]^{-1}. \tag{35}$$

If we set $\lambda_n = 1$, i.e. we want to decrease the KL between $\mu_{\lambda(t)}$ and π at a constant rate we obtain

$$\frac{d\lambda(t)}{dt} = c\left[(1-\lambda(t))\mathrm{I}(\lambda(t))\right]^{-1},$$

i.e. the ODE given in Goshtasbpour et al. (2023), where we used the fact that

$$\mathrm{Var}_{\mu_{\lambda(t)}}\left(\log\left(\frac{\pi}{\mu_{\lambda(t)}}(x)\right)\right) = (1-\lambda(t))^2\mathrm{Var}_{\mu_{\lambda(t)}}\left(\log\left(\frac{\pi}{\mu_0}(x)\right)\right) = (1-\lambda(t))^2\mathrm{I}(\lambda(t)). \tag{36}$$

If we instead assume that λ_n is sufficiently close to $\lambda(t)$ that $\lambda_n - \lambda(t) \approx {}^{d\lambda(t)}/_{dt}$ we obtain the ODE in (16). Or, equivalently, by discretizing (35) we obtain

$$\lambda_n - \lambda_{n-1} = c\left[(\lambda_n-\lambda_{n-1})\mathrm{I}(\lambda_{n-1})\right]^{-1},$$

which is equivalent to (15).

F.1. Numerical implementation for Figure 3

We reproduce the narrow Gaussian experiment of Goshtasbpour et al. (2023): the target is $\pi = \mathcal{N}(1_d, 0.1^2\,\mathrm{Id})$ and $\mu_0 = \mathcal{N}(0_d, \mathrm{Id})$ where $d = 2$.

To place all algorithms on equal footing we use the same number of particles $N = 10^4$ and the same Markov kernels, i.e. random-walk Metropolis kernels automatically calibrated on the current particle sample. In the case of SMC, we select the next tempering sequence so that $\mathrm{ESS}_n = N/2$, or, equivalently, by setting $\beta = 1$ in (14). For the constant rate AIS of (Goshtasbpour et al., 2023), we follow their recommendation and set $\delta = 1/32$ (higher values of δ give slightly shorter tempering sequences but considerably worse approximations of π). To make their algorithm more numerically stable we replace line 11 in their Algorithm 1 with (36). We also compare with SMC and AIS in which the step-size γ is constant $\gamma = 0.05$.

We point out that the resampling cost in SMC is negligible, and that a shorter tempering sequence does correspond to a shorter runtime (< 1 second for SMC and ≈ 14 seconds for AIS).