Error Bounds for Particle Gradient Descent,
and Extensions of the log-Sobolev and Talagrand Inequalities

Rocco Caprio   ROCCO.CAPRIO@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Juan Kuntz

Samuel Power   SAM.POWER@BRISTOL.AC.UK
School of Mathematics
University of Bristol
Bristol, BS8 1UG, UK

Adam M. Johansen   A.M.JOHANSEN@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Editor: Alexandre Bouchard

Abstract

We derive non-asymptotic error bounds for particle gradient descent (PGD, Kuntz et al. (2023)), a recently introduced algorithm for maximum likelihood estimation of large latent variable models obtained by discretizing a gradient flow of the free energy. We begin by showing that the flow converges exponentially fast to the free energy's minimizers for models satisfying a condition that generalizes both the log-Sobolev and the Polyak–Łojasiewicz inequalities (LSI and PŁI, respectively). We achieve this by extending a result well-known in the optimal transport literature (that the LSI implies the Talagrand inequality) and its counterpart in the optimization literature (that the PŁI implies the so-called quadratic growth condition), and applying the extension to our new setting. We also generalize the Bakry–Émery Theorem and show that the LSI/PŁI extension holds for models with strongly concave log-likelihoods. For such models, we further control PGD's discretization error and obtain the non-asymptotic error bounds. While we are motivated by the study of PGD, we believe that the inequalities and results we extend may be of independent interest.

Keywords: latent variable models, maximum marginal likelihood, gradient flows, log-Sobolev inequality, Polyak–Łojasiewicz inequality, Talagrand inequality, quadratic growth condition.

1. Introduction

Many tasks in machine learning and statistics require fitting a probabilistic model—with Lebesgue density, $p_\theta(x, y)$, and featuring latent variables, x—to data, y, we have observed. Often, we achieve this by finding model parameters, θ, that maximize the probability, $p_\theta(y)$,

of observing the data we observed (the *marginal likelihood*). That is, θ_* belonging to

$$\mathcal{O}_* := \arg\max_{\theta \in \mathbb{R}^{d_\theta}} p_\theta(y) = \arg\max_{\theta \in \mathbb{R}^{d_\theta}} \int p_\theta(x, y)\, \mathrm{d}x \qquad (1)$$

(we assume throughout that the latent variables and parameters respectively take values in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^{d_x}$). In many cases, the latent variables are meaningful or interesting in some way, and we would like to infer them. Following the empirical Bayes paradigm (Robbins, 1956), we do this using the posterior distribution of the latent variables given the data for the optimal parameters θ_*:

$$p_{\theta_*}(x|y) := \frac{p_{\theta_*}(x, y)}{p_{\theta_*}(y)}.$$

For most models of practical interest, the integral in (1) is intractable, we have no closed-form expressions for $p_\theta(y)$ or its derivatives, and we are unable to directly optimize $p_\theta(y)$. Often, this hurdle is overcome by noting that (θ_*, q_*) minimizes the *free energy* functional,

$$F(\theta, q) := \begin{cases} \int \log\left(\frac{q(x)}{p_\theta(x, y)}\right) q(\mathrm{d}x) & \text{if } q \ll \mathrm{d}x \\ +\infty & \text{otherwise} \end{cases} \qquad \forall (\theta, q) \in \mathcal{M}, \qquad (2)$$

if and only if θ_* maximizes the marginal likelihood and $q_* = p_{\theta_*}(\cdot \mid y)$; where $\mathrm{d}x$ in (2) denotes the Lebesgue measure and $\mathcal{M}$ the product of the parameter space, $\mathbb{R}^{d_\theta}$, and the space, $\mathcal{P}(\mathbb{R}^{d_x})$, of probability distributions over the latent space $\mathbb{R}^{d_x}$. For instance, the well-known expectation-maximization (EM) algorithm (Dempster et al., 1977) can be viewed as minimizing F using coordinate descent (CD) (Neal and Hinton, 1998). Similarly, many methods in variational inference (approximately) minimize F by (i) restricting $\mathcal{P}(\mathbb{R}^{d_x})$ to a parametric family of distributions $(q_\phi)_{\phi \in \Phi}$ such that the surrogate objective $(\theta, \phi) \mapsto F(\theta, q_\phi)$ is tractable, (ii) computing a minimizer (θ_*, ϕ_*) thereof using an appropriate optimization algorithm, and (iii) employing (θ_*, q_{ϕ_*}) as a proxy for a genuine minimizer of F; e.g., see Kingma and Welling (2019).

Inspired by the interpretation of the EM algorithm as CD applied to F, Kuntz et al. (2023) examines whether it is possible to minimize F over $\mathcal{M}$, rather than any restriction thereof, using analogues of optimization algorithms other than CD. In particular, they identify analogues of gradient descent (GD) applicable to F. To do so, they recall that the application of GD to minimizing a function $f : \mathbb{R}^{d_\theta} \to \mathbb{R}$ may be viewed as the forward Euler discretization of the ordinary differential equation (ODE) known as the *gradient flow*:

$$\dot{\theta}_t = -\nabla_\theta f(\theta_t), \qquad (3)$$

where ∇_θ denotes the usual Euclidean gradient on $\mathbb{R}^{d_\theta}$. Borrowing ideas from optimal transport, they identify $\nabla F = (\nabla_\theta F, \nabla_q F)$, with

$$\nabla_\theta F(\theta, q) = -\int \nabla_\theta \ell(\theta, x) q(\mathrm{d}x), \quad \nabla_q F(\theta, q) = \nabla_x \cdot \left[q \nabla_x \log\left(\frac{p_\theta(\cdot, y)}{q}\right)\right], \qquad (4)$$

and $\ell(\theta, x) := \log(p_\theta(x, y))$, as an analogue to $\nabla_\theta f$ and obtain the following gradient flow dynamics for the parameter estimate θ_t and the latent posterior approximation q_t:

$$\dot{\theta}_t = \int \nabla_\theta \ell(\theta_t, x) q_t(\mathrm{d}x), \qquad \dot{q}_t = \nabla_x \cdot \left[q_t \nabla_x \log \left(\frac{q_t}{p_{\theta_t}(\cdot|y)} \right) \right]. \tag{5}$$

They go on to (i) interpret (5) as the Fokker–Planck equation for the following McKean–Vlasov stochastic differential equation (SDE; e.g., see Chaintron and Diez (2022)):

$$\mathrm{d}\theta_t = \int \nabla_\theta \ell(\theta_t, x) q_t(\mathrm{d}x)\, \mathrm{d}t, \quad \mathrm{d}X_t = \nabla_x \ell(\theta_t, X_t)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t, \text{ where } q_t := \mathrm{Law}(X_t); \tag{6}$$

and (ii) to approximate the law q_t with an empirical distribution $N^{-1} \sum_{n=1}^N \delta_{X^n}$ of *particles* $X^1, \ldots, X^N$; and (iii) discretize the resulting SDE in time. The result is a practical algorithm for fitting latent variable models they refer to as particle gradient descent (PGD, Algorithm 1; see also Wang et al. (2025) for a version suited to latent diffusion models):

Algorithm 1 Particle gradient descent (PGD).

1: **Inputs:** step size h, step number K, particle number N, and initial particles $X_0^{1,h}, \ldots,$ $X_0^{N,h}$ and parameter $\Theta_0^{N,h}$.

2: **for** $k = 0, \ldots, K-1$ **do**

3:    Update the parameter estimate:
$$\Theta_{k+1}^{N,h} = \Theta_k^{N,h} + \frac{h}{N} \sum_{n=1}^N \nabla_\theta \ell(\Theta_k^{N,h}, X_k^{n,h}). \tag{7}$$

4:    Update the particles: with $W_k^1, \ldots, W_k^N$ denoting i.i.d. $\mathcal{N}(0, I_{d_x})$ r.v.s,
$$X_{k+1}^{n,h} = X_k^{n,h} + h \nabla_x \ell(\Theta_k^{N,h}, X_k^{n,h}) + \sqrt{2h} W_k^n \quad \forall n = 1, \ldots, N. \tag{8}$$

5: **end for**

6: **return** $\Theta_K^{N,h}$ and $Q_K^{N,h} := N^{-1} \sum_{n=1}^N \delta_{X_K^{n,h}}$.

PGD performs well in experiments (Kuntz et al., 2023) and is well-suited to modern computing environments. Its updates (7,8) require only evaluating $\nabla\ell := (\nabla_\theta \ell, \nabla_x \ell)$ and adding noise, thus facilitating its implementation in modern autodiff frameworks such as TensorFlow, PyTorch, and JAX. Moreover, each particle's update (8) is independent of all other particles, offering scope for simple parallelisation and distribution of these steps. Implemented naively, a full run of the algorithm costs $\mathcal{O}(NK[\text{evaluation cost of } \nabla\ell])$ operations, where N denotes the number of particles employed and K the total number of steps taken. The running time's dependence on N can be further reduced through the aforementioned parallelism strategies.

Here, we validate theoretically PGD and show that, for models whose log-likelihoods ℓ are λ-strongly concave with L-Lipschitz gradients, its output, $(\Theta_K^{N,h}, Q_K^{N,h})$, satisfies

$$\mathsf{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\theta_*, Q_*^N)) = \mathcal{O}(h^{\frac{1}{2}} + N^{-\frac{1}{2}} + e^{-h\lambda K}), \tag{9}$$

assuming that the algorithm's step size h is no greater than $1/(\lambda + L)$. In the above, Q_*^N denotes the empirical distribution of N i.i.d. particles drawn from $p_{\theta_*}(\cdot|y)$ and d the

following metric acting on $\mathcal{M}$-valued random variables:

$$\mathsf{d}((\Theta, Q), (\Theta', Q')) := \sqrt{\mathbb{E}\left[\mathsf{d}_E(\Theta, \Theta')^2\right] + \mathbb{E}\left[\mathsf{d}_{W_2}(Q, Q')^2\right]} \tag{10}$$

where d_E and d_{W_2} respectively denote the Euclidean and Wasserstein-2 metrics.

We begin by studying PGD's continuous-time infinite-particle limit: the gradient flow (5). First, we show that the free energy $(F(\theta_t, q_t))_{t \geq 0}$ evaluated along the flow converges at an exponential rate to its infimum,

$$F_* := \inf_{(\theta, q) \in \mathcal{M}} F(\theta, q),$$

provided that the model satisfies a condition which generalizes the log-Sobolev inequality (LSI) popular in optimal transport (e.g., see Villani, 2009, Chapter 21) and the Polyak–Łojasiewicz inequality (PŁI) used in optimization (Karimi et al., 2016). In particular, we assume that there exists a constant $\lambda > 0$ such that

$$2\lambda[F(\theta, q) - F_*] \leq I(\theta, q) \tag{11}$$

for all parameters θ and distributions q, where the functional I is defined by

$$I(\theta, q) := \left\| \int \nabla_\theta \ell(\theta, x) q(\mathrm{d}x) \right\|^2 + \int \left\| \nabla_x \log\left(\frac{q(x)}{p_\theta(x, y)} \right) \right\|^2 q(\mathrm{d}x), \tag{12}$$

with $\|\cdot\|$ denoting the Euclidean norm. We then extend a result of Otto and Villani (2000) and show that models which satisfy (11) also satisfy the following generalization of both the Talagrand inequality (Talagrand, 1996) and the *quadratic growth* condition used in optimization (Anitescu, 2000; Karimi et al., 2016): for all parameters θ and distributions q, it holds that

$$2[F(\theta, q) - F_*] \geq \lambda \mathsf{d}((\theta, q), \mathcal{M}_*)^2, \tag{13}$$

where $\mathcal{M}_*$ denotes F's optimal set and $\mathsf{d}((\theta, q), \mathcal{M}_*)$ the distance from (θ, q) to $\mathcal{M}_*$:

$$\mathcal{M}_* := \underset{(\theta, q) \in \mathcal{M}}{\arg\min} F(\theta, q), \quad \mathsf{d}((\theta, q), \mathcal{M}_*) := \inf_{(\theta_*, q_*) \in \mathcal{M}_*} \mathsf{d}((\theta, q), (\theta_*, q_*)). \tag{14}$$

Using (13), we then show that the flow itself $(\theta_t, q_t)_{t \geq 0}$ converges to $\mathcal{M}_*$ exponentially fast in d. Next, we generalize the Bakry–Émery Theorem and show that models with strongly concave likelihoods ℓ satisfy (11). For such models, we obtain a slightly stronger convergence result. Under the further assumption that ℓ's gradients are Lipschitz, we are able to bound the errors introduced by both approximating distributions with empirical measures and discretizing (5) in time, and the bound (9) follows.

1.1 Related Literature

The analysis provided in Kuntz et al. (2023) was limited: as explained in the 'Our setting, notation, assumptions, rigour, and lack thereof' paragraph of the paper's introduction, it focused on intuition and practical application rather than theoretical validation. While the authors obtained no error bounds, they did argue with their Theorem 3 that the gradient

flow (5) converges exponentially fast to F's minimizer whenever ℓ is strongly concave. However, they imposed an unnecessary technical condition that substantially restricts the applicability of their result (that the gradient of ℓ be bounded uniformly).

Recently, several works have proposed particle-based algorithms for maximum marginal likelihood estimation. Akyildiz et al. (2025) consider an alternative approximation to the flow (5) which they term the 'Interacting Particle Langevin Algorithm' or IPLA; see also Encinar et al. (2024) for a version applicable to non-differentiable models. IPLA differs from PGD by the inclusion of an extra additive Gaussian noise term $\sqrt{2h/N}W_k$ in the parameter's update (7). This difference enables an analysis of the algorithm using tools similar to those used to study the unadjusted Langevin algorithm in Durmus and Moulines (2017, 2019). In particular, the authors show that the $\mathcal{L}^2$ error of the parameter estimates produced by IPLA satisfies (9,RHS) for models satisfying the same Lipschitz gradient and strong concavity assumptions made here, which we find pleasantly consistent with our results in Section 3. Lim et al. (2024); Chen (2023); Oliva and Akyildiz (2024) consider variants of PGD and IPLA that incorporate momentum in the updates and can outperform PGD and IPLA in practice. These variants no longer approximate the gradient flow (5) and, consequently, require a different type of analysis. In particular, Lim et al. (2024) established exponentially fast convergence of their algorithm's limiting (non-gradient) flow under (11), while Oliva and Akyildiz (2024) obtained non-asymptotic error bounds for their algorithms' parameter estimates also under Lipschitz gradient and strong concavity assumptions. Sharrock et al. (2024) introduced two new particle-based methods for minimizing F. The first, 'Stein Variational Gradient Descent EM' or SVGD EM, is also constructed as an approximation to a gradient flow of F, but with respect to a different underlying geometry—the so-called 'Stein Geometry' (Liu and Wang, 2016; Duncan et al., 2023), as opposed to the Euclidean/Wasserstein-2 geometry underpinning (4) (whose induced distance metric is given by (10) omitting the expectations; cf. Kuntz et al., 2023, Appendix A for more on this). The second, 'Coin EM', involves an entirely different approach for minimizing F that builds on coin betting techniques from convex optimization and obviates the need for tuning discretization step sizes. Only SVGD-EM is analyzed theoretically. To do so, the authors assume a natural 'Stein' analogue of (11): Assumption 7 in Sharrock et al. (2024, Appendix A). Fan et al. (2023) considers a non-parametric maximum likelihood estimation problem where they optimize a measure over the parameters rather than the parameters themselves by approximating F's gradient flow in the Fisher–Rao/Wasserstein-2 geometry. Finally, Crucinio (2025) recently proposed an approach for maximum marginal likelihood estimation based on mirror descent and sequential Monte Carlo methods to minimize F.

Our results in Section 2 relate to earlier results well-known in the optimal transport and optimization literatures. We discuss these connections as we go along.

1.2 Paper Structure

In Section 2, we study the convergence of the gradient flow (5) and the inequalities (11,13). In Section 3, we analyze the flow's approximation (PGD; Algorithm 1) and obtain non-asymptotic bounds on its error. We conclude in Section 4 by discussing our results beyond the context of PGD.

2. Convergence of the Gradient Flow

In this section, we study the gradient flow (5) and several pertinent inequalities. As we explain in what follows, these inequalities generalize others well-known in the literature, and we prefix the former with 'extended' to differentiate them from the latter. We start by showing that for models satisfying the extended LSI (xLSI) (11), the values of F along the flow converge exponentially fast to the infimum of F (Section 2.1). We next argue that any model satisfying the xLSI also satisfies the extended Talagrand-type inequality (13), and we use this result to show that the trajectory of the flow itself converges exponentially fast to $(\theta_*, p_{\theta_*}(\cdot|y))$ for some maximizer θ_* of the marginal likelihood (Section 2.2). Finally, we show that the xLSI holds for models with λ-strongly concave log-likelihoods ℓ, and we obtain a slightly stronger convergence result for this case (Section 2.3).

Notation and Assumptions. For the remainder of the paper, we write $\rho_\theta(\cdot)$ for the likelihood $p_\theta(\cdot, y)$, Z_θ for the marginal likelihood $p_\theta(y)$, and $\pi_\theta(\cdot)$ for the posterior distribution $p_\theta(\cdot \mid y)$. To ensure that d in (10) is a metric, we consider the restriction of $\mathcal{P}(\mathbb{R}^{d_x})$ to the subset of measures with finite second moments, and similarly for $\mathcal{M}$:

$$\mathcal{P}_2(\mathbb{R}^{d_x}) := \left\{ q \in \mathcal{P}(\mathbb{R}^{d_x}) : \int \|x\|^2 \, q(\mathrm{d}x) < \infty \right\}, \quad \mathcal{M}_2 := \mathbb{R}^{d_\theta} \times \mathcal{P}_2(\mathbb{R}^{d_x}).$$

To ensure that the functional I in (12) is well-defined, we further restrict restrict $\mathcal{P}_2(\mathbb{R}^{d_x})$ to the subset of measures with densities differentiable almost everywhere w.r.t. to the Lebesgue measure $\mathrm{d}x$, and similarly for $\mathcal{M}_2$:

$$\mathcal{P}_2^1(\mathbb{R}^{d_x}) := \left\{ q \in \mathcal{P}_2(\mathbb{R}^{d_x}) : q \ll \mathrm{d}x, \, \nabla_x \frac{\mathrm{d}q}{\mathrm{d}x}(x) \text{ exists a.e.} \right\}, \quad \mathcal{M}_2^1 := \mathbb{R}^{d_\theta} \times \mathcal{P}_2^1(\mathbb{R}^{d_x}).$$

We also assume that the model is suitably differentiable:

Assumption 1 (Model regularity) *(i) For all x in $\mathbb{R}^{d_x}$, $\theta \mapsto \pi_\theta(x)$ is differentiable; and $\theta \mapsto Z_\theta$ is differentiable; (ii) for all θ in $\mathbb{R}^{d_\theta}$, π_θ is twice continuously differentiable; (iii) for all θ in $\mathbb{R}^{d_\theta}$ and x in $\mathbb{R}^{d_x}$, $\rho_\theta(x) > 0$;*

Next, we assume that the gradient flow (5) has sufficiently regular solutions. To this end, let $\mathcal{C}^i(\mathcal{X}, \mathcal{Y})$ denote the space of i-times continuously differentiable functions from $\mathcal{X}$ to $\mathcal{Y}$, $\mathcal{C}_c^i(\mathcal{X}, \mathcal{Y})$ the subspace of such functions with compact support, and $\mathcal{C}^{i,j}(\mathcal{X} \times \mathcal{X}', \mathcal{Y})$ denote the functions that are $\mathcal{C}^i(\mathcal{X}, \mathcal{Y})$ in the first variable and $\mathcal{C}^j(\mathcal{X}', \mathcal{Y})$ in the second.

Assumption 2 (Regularity of solutions) *For any initial conditions (θ, q) in $\mathcal{M}_2$, (5) has a classical solution $(\theta_t, q_t)_{t \geq 0}$ with $(\theta_0, q_0) = (\theta, q)$. For any such solution, $q_t(\mathrm{d}x)$ has a Lebesgue density $q_t(x)$ for all $t > 0$; $(t, x) \mapsto q_t(x)$ belongs to $\mathcal{C}^{1,2}([0, \infty) \times \mathbb{R}^{d_x}, (0, \infty))$, and $t \mapsto \theta_t$ belongs to $\mathcal{C}^1([0, \infty), \mathbb{R}^{d_\theta})$.*

We verify in Section 3 that the above holds whenever the model's log-likelihood ℓ has a Lipschitz gradient. We believe it will also hold under weaker conditions on ℓ because, in the special case that π_θ is a single distribution π independent of θ, the flow (5) essentially reduces to the following well-known Fokker–Planck equation,

$$\dot{q}_t = \nabla_x \cdot \left[q_t \nabla_x \log\left(\frac{q_t}{\pi} \right) \right]; \tag{15}$$

and stronger regularity properties for (15) have been established under weaker conditions on ℓ. For example, Jordan et al. (1998, Theorem 5.1) shows that the solutions are smooth whenever $\log(\pi)$ is smooth.

2.1 The Extended log-Sobolev Inequality and $(F(\theta_t, q_t))_{t \geq 0}$'s Convergence

We prove the flow's convergence for models which satisfy the xLSI in the following sense:

Definition 1 (Extended log-Sobolev inequality (xLSI)) *We say that the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the extended log-Sobolev inequality (xLSI) with constant $\lambda > 0$ if (11) holds for all (θ, q) in $\mathcal{M}_2^1$.*

Under this assumption, the values of F converge along the gradient flow exponentially fast to F's infimum over $\mathcal{M}_2$:

Theorem 2 (xLSI $\Rightarrow$ exp. conv. of $(F(\theta_t, q_t))_{t \geq 0}$) *If Assumptions 1–2 hold, (θ_0, q_0) belongs to $\mathcal{M}_2$, and the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the xLSI with constant $\lambda > 0$, then*

$$0 \leq F(\theta_t, q_t) - F_* \leq [F(\theta_0, q_0) - F_*]e^{-2\lambda t} \quad \forall t \geq 0. \tag{16}$$

Proof. The leftmost inequality in (16) follows from the F_*'s definition. As we show in Lemma 15 in Appendix A.1, the following extension to de Bruijn's identity holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\theta_t, q_t) = -I(\theta_t, q_t) \quad \forall t > 0. \tag{17}$$

Now the rightmost equality follows by combining the xLSI (11) with (17) and applying Grönwall's inequality. ∎

As is well known,
$$F(\theta, q) = \mathrm{KL}(q||\pi_\theta) - \log(Z_\theta) \quad \forall (\theta, q) \in \mathcal{M}, \tag{18}$$

where KL denotes the Kullback–Leibler divergence:

$$\mathrm{KL}(q||\pi) := \begin{cases} \displaystyle\int \log\left(\frac{q(x)}{\pi(x)}\right) q(\mathrm{d}x) & \text{if } q \ll \mathrm{d}x \\ +\infty & \text{otherwise} \end{cases} \quad \forall q \in \mathcal{P}(\mathbb{R}^{d_x}),$$

Because $\mathrm{KL}(q||q')$ is non-negative and zero iff $q = q'$ a.e., (18) implies that

$$F_* = \inf_{(\theta, q) \in \mathcal{M}} F(\theta, q) = \inf_{\theta \in \mathbb{R}^{d_\theta}} F(\theta, \pi_\theta) = -\log\left(\sup_{\theta \in \mathbb{R}^{d_\theta}} Z_\theta\right) = -\log Z_*, \tag{19}$$

where $Z_* := \sup_{\theta \in \mathbb{R}^{d_\theta}} Z_\theta$ denotes the marginal likelihood's supremum. Putting (16–19) together, we find that

$$\log(Z_*) - \log(Z_{\theta_t}) + \mathrm{KL}(q_t||\pi_{\theta_t}) = F(\theta_t, q_t) + \log(Z_*) = \mathcal{O}(e^{-2\lambda t}).$$

In other words, Theorem 2 shows that as t increases: the free energy $F(\theta_t, q_t)$ converges exponentially fast to $-\log Z_*$, the log-marginal likelihood $\log(Z_{\theta_t})$ converges exponentially

fast to its supremum $\log(Z_*)$, and q_t tracks the corresponding posterior distributions π_{θ_t} in the sense that the KL divergence between the two decays exponentially fast to zero.

Connections with the log-Sobolev inequality. We refer to the inequality in Definition 1 as the 'xLSI' because it extends the classical log-Sobolev inequality (LSI) (Gross, 1975); e.g., see Villani (2009, Definition 21.1). While the xLSI is a statement about a parametrized family of measures, the LSI is a statement about a single probability measure. In particular, if π_θ is a distribution π independent of θ, then (11) reads

$$2\lambda \mathrm{KL}(q||\pi) \le I(q||\pi) \ \ \forall q \in \mathcal{P}_2^1(\mathbb{R}^{d_x}); \ \ \text{where} \ \ I(q||\pi) := \int \left\| \nabla_x \log\left(\frac{q(x)}{\pi(x)}\right) \right\|^2 q(\mathrm{d}x). \quad (20)$$

For this reason, Theorem 2 extends a well-known result (e.g., Bakry et al. (2014, Section 5.2)) stating that, if π satisfies the LSI (20) and $(q_t)_{t\ge0}$ solves the Fokker–Planck equation (15), then the KL divergence from q_t to π decays exponentially fast.

Connections with the Polyak–Łojasiewicz inequality. Definition 1 also extends an inequality due to Polyak (1963) and Łojasiewicz (1963) commonly used to argue linear convergence of gradient descent algorithms. A differentiable real-valued function f on $\mathbb{R}^{d_\theta}$ with infimum f_* is said to satisfy the PLI with constant $\lambda > 0$ if

$$2\lambda[f(\theta) - f_*] \le \left\| \nabla_\theta f(\theta) \right\|^2 \quad \theta \in \mathbb{R}^{d_\theta}. \quad (21)$$

In particular, because

$$\int \nabla_\theta \ell(\theta, x)\pi_\theta(\mathrm{d}x) = \int \frac{\nabla_\theta \rho_\theta(x)}{\rho_\theta(x)}\pi_\theta(\mathrm{d}x) = \frac{\nabla_\theta \int \rho_\theta(x)\,\mathrm{d}x}{Z_\theta} = \frac{\nabla_\theta Z_\theta}{Z_\theta} = \nabla_\theta \log(Z_\theta), \quad (22)$$

setting $q := \pi_\theta$ in (11), we recover (21) with $f(\theta) := -\log(Z_\theta)$ (the exchange of limits in (22) is valid whenever Fisher's identity holds, which is a common assumption in the study of latent variable models and the EM algorithm; it is satisfied whenever e.g. π_θ belongs to an exponential family, see Douc et al. 2014, Appendix D). In this case, Theorem 2 reduces to a well-known result showing that f converges exponentially fast along the Euclidean gradient flow (3) to its infimum whenever f satisfies the PLI; e.g., see Trillos et al. (2023, Proposition 2.3). Among other reasons, the PLI (21) is popular in optimization because—without requiring the objective to be convex—it implies that the objective's stationary points are global minimizers. In our case, the xLSI implies that the marginal likelihood's stationary points are global maximizers.

2.2 The Talagrand Inequality and the Flow's Exponential Convergence

We now show that, for models satisfying the xLSI, the gradient flow $(\theta_t, q_t)_{t\ge0}$ itself converges exponentially fast to the free energy's minimizers. We measure the convergence in terms of the following metric on $\mathcal{M}_2$:

$$\mathsf{d}((\theta, q), (\theta', q')) := \sqrt{\mathsf{d}_E(\theta, \theta')^2 + \mathsf{d}_{W_2}(q, q')^2}. \quad (23)$$

Since F is a lower semicontinuous function on $(\mathcal{M}_2, \mathsf{d})$ (see Lemma 18 in Appendix A.2), the set $\mathcal{M}_*$ in (14) of the free energy's minima is closed (note that Assumption 1 and (19) imply that $\mathcal{M}_*$ is contained in $\mathcal{M}_2$). Using (18,19), we can characterize $\mathcal{M}_*$ in terms of the marginal likelihood's set of maximizers ($\mathcal{O}_*$ in (1)):

$$\mathcal{M}_* = \{(\theta_*, \pi_{\theta_*}) : \theta_* \in \mathcal{O}_*\}. \tag{24}$$

To argue the flow's convergence, we require the following extension of the Talagrand inequality:

Definition 3 (Extension of the Talagrand inequality (xT$_2$I)) *We say that the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy an extension of the Talagrand inequality (xT$_2$I) with constant $\lambda > 0$ if (13) holds for all (θ, q) in $\mathcal{M}_2$.*

The inequality holds for all models satisfying the xLSI:

Theorem 4 (xLSI $\Rightarrow$ xT$_2$I) *If Assumptions 1–2 hold, and the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the xLSI with constant $\lambda > 0$, then they also satisfy the xT$_2$I with the same constant.*

See Appendix A.2 for the proof. The convergence of $(\theta_t, q_t)_{t \geq 0}$ then follows from Theorem 2:

Corollary 5 (xLSI $\Rightarrow$ exp. conv. of $(\theta_t, q_t)_{t \geq 0}$) *If Assumptions 1–2 hold, (θ_0, q_0) belongs to $\mathcal{M}_2$, and the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the xLSI with constant $\lambda > 0$, then*

$$\lambda \mathsf{d}((\theta_t, q_t), \mathcal{M}_*)^2 \leq 2[F(\theta_0, q_0) - F_*]e^{-2\lambda t} \quad \forall t \geq 0.$$

Connections with the Talagrand inequality. By (24), if π_θ is a distribution π independent of θ, then $\mathcal{M}_* = \mathbb{R}^{d_\theta} \times \{\pi\}$ and (13) reduces to the Talagrand inequality (Talagrand, 1996, Theorem 1.2):

$$2\mathrm{KL}(q||\pi) \geq \lambda \mathsf{d}_{W_2}(q, \pi)^2 \quad \forall q \in \mathcal{P}_2(\mathbb{R}^{d_x}). \tag{25}$$

For this reason, Theorem 4 extends the Otto–Villani Theorem (Otto and Villani, 2000, Theorem 1) showing that the LSI (20) implies (25).

Connections with the quadratic growth condition. Setting $(q, f(\theta), f_*)$ to $(\pi_\theta, -\log(Z_\theta), F_*)$, (13) reduces to the 'quadratic growth' condition used in optimization to establish linear convergence rates to local minima of gradient descent algorithms:

$$2[f(\theta) - f_*] \geq \lambda \mathsf{d}_E(\theta, \mathcal{O}_*)^2 \quad \forall \theta \in \mathbb{R}^{d_\theta}. \tag{26}$$

In this case, Theorem 4 reduces to the well-known result stating that (26) holds whenever the PŁI (21) holds; see Karimi et al. (2016, Theorem 2).

2.3 Strongly log-Concave Models

In Section 3, we study PGD's convergence for models satisfying the following assumption.

Assumption 3 (Strong log-concavity) *There exists a $\lambda > 0$ such that*

$$\ell((1-t)\theta + t\theta', (1-t)x + tx') \geq (1-t)\ell(\theta, x) + t\ell(\theta', x') + \frac{\lambda t(1-t)}{2}\left\|(\theta, x) - (\theta', x')\right\|^2,$$

for all $(\theta, x), (\theta', x')$ in $\mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}$ and $0 \leq t \leq 1$.

Models that satisfy the above also satisfy the xLSI:

Theorem 6 (Strong log-concavity $\Rightarrow$ xLSI) *Any model satisfying Assumptions 1 and 3 satisfies the xLSI with constant $\lambda > 0$.*

Proof. The result follows from the fact that the free energy is strongly geodesically convex if log-likelihood is strongly concave; see Appendix A.3 for details. ∎

Simple examples satisfying Assumption 3, and consequently the xLSI, include Bayesian logistic regression (e.g., Genkin et al., 2007; see the proof of Proposition 1 in Appendix E.2 of Kuntz et al., 2023 for an argument) and certain types of hierarchical models (Gelman and Hill, 2007; see Caprio and Johansen, 2024, Example 1 for details).

Connection with the Bakry–Émery Theorem. Given that (11) reduces to the standard LSI (20) whenever π_θ is independent of θ, Theorem 6 extends Bakry–Émery Theorem (Bakry and Émery, 1985) showing that distributions with strongly log-concave densities satisfy (20).

Connection with the optimization literature. Given that (11) with $q = \pi_\theta$ reduces to the PŁI (21) with $f(\theta) = -\log(Z_\theta)$, Theorem 6 extends the well-known fact that the PŁI holds for strongly convex f; e.g., see Karimi et al. (2016, Theorem 2).

The flow's convergence. Under Assumption 3, the marginal likelihood has a unique maximizer θ_*; e.g., see Kuntz et al. (2023, Theorem 4 in Appendix B.3). Hence, Corollary 5 and (24) tell us that the flow converges exponentially fast to $(\theta_*, \pi_{\theta_*})$:

$$\lambda \mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*}))^2 \leq 2[F(\theta_0, q_0) - F_*]e^{-2\lambda t} \quad \forall t \geq 0.$$

In this special case, we can sharpen the above slightly (see Appendix A.4 for the proof):

Theorem 7 (Strong log-concavity $\Rightarrow$ exp. conv. of $(\theta_t, q_t)_{t\geq0}$) *If Assumptions 1(iii), 2, and 3 hold, and (θ_0, q_0) belongs to $\mathcal{M}_2$, then*

$$\mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*})) \leq \mathsf{d}((\theta_0, q_0), (\theta_*, \pi_{\theta_*}))e^{-\lambda t} \quad \forall t \geq 0.$$

3. Error Bounds for Particle Gradient Descent

Here, we capitalize on the results of Section 2 to obtain the error bound in (9) for PGD (Algorithm 1) applied to models with strongly concave log-likelihoods. In particular, Theorem 7 in Section 2.3 bounds the distance between the gradient flow and $(\theta_*, \pi_{\theta_*})$, where θ_* denotes the marginal likelihood's unique maximizer and the distance is measured in terms of the metric d in (23). To exploit this bound and obtain (9), we first need to connect the gradient flow to PGD, which we do via the McKean–Vlasov SDE (6). To ensure that the SDE has globally-defined solutions, we assume that ℓ's gradient is Lipschitz:

Assumption 4 *The log-likelihood ℓ is differentiable and its gradient $\nabla\ell = (\nabla_\theta\ell, \nabla_x\ell)$ is L-Lipschitz for some $L > 0$:*

$$\left\|\nabla\ell(\theta, x) - \nabla\ell(\theta', x')\right\| \leq L\left\|(\theta, x) - (\theta', x')\right\| \quad \forall (\theta, x), (\theta', x') \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}.$$

Let $\mathcal{L}^2(\mathbb{R}^{d_x})$ denote the space of square-integrable random variables taking values in $\mathbb{R}^{d_x}$. The SDE's relation to the gradient flow is as follows (see Appendix B.1 for the proof):

Proposition 8 *If Assumptions 1(ii) and 4 hold and (θ, X) belongs to $\mathbb{R}^{d_\theta} \times \mathcal{L}^2(\mathbb{R}^{d_x})$, then the SDE (6) has an unique strong solution $(\theta_t, X_t)_{t \geq 0}$ such that $(\theta_0, X_0) = (\theta, X)$. Moreover, $(\theta_t, \mathrm{Law}(X_t))_{t \geq 0}$ is a classical solution of the gradient flow (5) satisfying Assumption 2.*

PGD, on the other hand, is obtained by approximating (6) as we detail in Section 3.3 below. Bounding the errors introduced by these approximations, we obtain (9). In particular, recall that $(\Theta_K^{N,h}, Q_K^{N,h})$ denotes PGD's output after K iterations and Q_*^N the empirical distribution of N i.i.d. particles drawn from π_{θ_*}. Moreover, note that, the log-likelihood ℓ has a unique maximizer $(\theta_\dagger, x_\dagger)$ which, without loss of generality, we assume lies at the origin $(0, 0)$. Our main result for PGD is then as follows (see Section 3.3 for the proof):

Theorem 9 (PGD error bound) *Under Assumptions 1(ii–iii) and 3–4, if $X_0^{1,h}, \ldots, X_0^{N,h}$ are drawn independently from a distribution q_0 in $\mathcal{P}_2(\mathbb{R}^{d_x})$ and $h \leq 1/(\lambda + L)$, then*

$$\mathsf{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\theta_*, Q_*^N)) \leq \sqrt{h} A_{0,h} + \frac{L\sqrt{2}}{\lambda\sqrt{N}} \sqrt{B_0 + \frac{2d_x}{\lambda}} + \mathsf{d}((\theta_0, q_0), (\theta_*, \pi_{\theta_*})) e^{-h\lambda K} \quad (27)$$

for all K in $\mathbb{N}$; where $B_0 := \|\theta_0\|^2 + \mathbb{E}\left[\|X_0\|^2\right]$ and

$$A_{0,h} := \sqrt{\frac{4h + 4/\iota}{\iota} 220 L^2 \left(L^2 h \left[B_0 + \frac{2d_x}{\lambda}\right] + d_x\right)} \quad with \quad \iota := \frac{2L\lambda}{L + \lambda}.$$

In short, Theorem 9 gives us the bound (9) and an explicit expression for the proportionality constant as a function of the algorithm's parameters, the model's Lipschitz and concavity constants, the initial condition, and the latent space's dimensionality. It follows that in order to achieve an error of $\mathcal{O}(\epsilon)$, we need only to choose the step size h proportional to ϵ^2, the particle number N proportional to ϵ^{-2}, and the step number K proportional to $\epsilon^{-2} \log(\epsilon^{-1})$. The corresponding computational cost is $\mathcal{O}(\epsilon^{-4} \log(\epsilon^{-1})[\text{eval. cost of } \ell])$; and we can lower the running time to $\mathcal{O}(\epsilon^{-2} \log(\epsilon^{-1})[\text{eval. cost of } \ell])$ by parallelizing/distributing computations across the particles (recall that each particle's update is independent of the other particles').

Given the definition of the metric d in (10), Theorem 9 bounds the $\mathcal{L}^2$ error of the parameter estimates, $\Theta_K^{N,h}$, and, also, the expected W_2 distance between the particles' empirical distribution, $Q_K^{N,h}$, and that, Q_*^N, of N i.i.d. samples drawn from the optimal posterior π_{θ_*}. We find the latter insightful because, in many ways, Q_*^N is the best one could hope to achieve using an $\mathcal{O}(N)$-cost Monte Carlo algorithm that returns N-sample approximations π_{θ_*}. The expected W_2 distance between Q_*^N itself and the optimal posterior π_{θ_*} is essentially a property of π_{θ_*} and has more to do with the fundamental limitations of approximating π_{θ_*} using N-sample ensembles than the method used to obtain the ensemble. To bound this distance (and, by combining this bound with Theorem 9, to obtain bounds on the distance between $Q_K^{N,h}$ and π_{θ_*}), we point the reader to Fournier (2023), Dereich et al. (2013), and references therein.

3.1 The error bound under warm-starts

If we 'warm start' PGD (e.g., see Dalalyan, 2017) by initializing the parameter estimates and the particles at the log-likelihood ℓ maxima (i.e., with $\theta_0 = 0$ and $q_0 = \delta_0$), then B_0 in (27) vanishes and we obtain the following sharper bound.

Corollary 10 (Error bound under warm starts) *Under the conditions in Theorem 9's premise and with ι as defined therein, if $(\theta_0, q_0) := (0, \delta_0)$, then, for all K in $\mathbb{N}$,*

$$\mathsf{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\theta_*, Q_*^N)) \leq \sqrt{h\frac{4h + 4/\iota}{\iota}220L^2\left(\frac{2L^2hd_x}{\lambda} + d_x\right) + 2\sqrt{\frac{d_x}{\lambda}}\left(\frac{L}{\lambda\sqrt{N}} + e^{-h\lambda K}\right)}.$$

Proof. Combining Theorem 7 together with Proposition 26 in Appendix B.1, we find that

$$\mathsf{d}((\theta_0, q_0), (\theta_*, \pi_{\theta_*}))^2 = \lim_{t\to\infty}\mathsf{d}((\theta_0, q_0), (\theta_t, q_t))^2 \leq \|\theta_t\|^2 + \mathbb{E}\left[\|X_t\|^2\right] \leq \frac{4d_x}{\lambda}.$$

∎

This bound features only PGD's parameters, the model's Lipschitz and concavity constants, and the latent space's dimensionality; and we can be more explicit in our error analysis. To measure the error, we use $\sqrt{\lambda}\mathsf{d}$ rather than d: a common practice in the analysis of Langevin algorithms (e.g., see Chewi, 2024, Section 4.1) motivated by $\sqrt{\lambda}\mathsf{d}$'s close relationship with the objectives we wish to minimize (13,25,26). In this metric, and with $\kappa := L/\lambda$ denoting the model's condition number, Corollary 10 shows that PGD, with a warm start, achieves an $\mathcal{O}(\epsilon)$ error as long as the step size satisfies $h \asymp \epsilon^2(d_x L\kappa)^{-1}$, the number of particles is $\mathcal{O}(d_x\kappa^2\epsilon^{-2})$, and the number of time steps is $\mathcal{O}(d_x\kappa^2\log(d_x^{1/2}\epsilon^{-1})\epsilon^{-2})$. The corresponding cost is $\mathcal{O}(d_x^2\kappa^4\log(d_x^{1/2}\epsilon^{-1})\epsilon^{-4}[\text{eval. cost of }\ell])$ and the running time can be brought down to $\mathcal{O}(d_x\kappa^2\log(d_x^{1/2}\epsilon^{-1})\epsilon^{-2}[\text{eval. cost of }\ell])$ by parallelizing/distributing over particles. Starting PGD warm is reasonably practical because the cost of running PGD will typically significantly exceed that of maximizing ℓ using a first-order method.

The cost and running time bounds grow with the latent space's dimensionality d_x. This is an issue for many models of practical interest where d_x is proportional to the number of datapoints in a large training set. However, as we now explore, it is possible to obtain bounds independent of d_x for a class of such models.

3.2 Dimension-free bounds for models with independent latent variables and conditionally Gaussian observations

Often, in machine learning applications, the data y decomposes as a sequence of datapoints $(\tilde{y}_m)_{m=1}^M$, the model features a latent variable $\tilde{x}_m$ per datapoint $\tilde{y}_m$, the pairs $(\tilde{x}_m, \tilde{y}_m)_{m=1}^M$ are assumed to be i.i.d. and generated through a mechanism of the sort

$$\tilde{y}_m = f_\theta(\tilde{x}_m) + \eta_m, \quad \tilde{x}_m \sim p_\theta^x(\mathrm{d}\tilde{x}), \quad \eta_m \sim p^\eta(\mathrm{d}\eta), \quad \forall m = 1, \ldots, M;$$

where f_θ denotes a parametrized function mapping from a latent space $\mathbb{R}^{\tilde{d}_x}$ to the parameter space, p_θ^x denotes a prior distribution on the latent variables, and the noise has a Gaussian law, p^η, which is independent of the model parameters θ. Examples include noisy independent component analysis (Hyvärinen, 1998), probabilistic matrix factorization (Mnih and

Salakhutdinov, 2007), the generative model behind variational autoencoders (Kingma and Welling, 2019), and latent diffusion models (Vahdat et al., 2021; Wehenkel and Louppe, 2021; Wang et al., 2025). In these cases, the log-likelihood breaks down into a sum of per-datapoint log-likelihoods:

$$\ell(\theta, x) = \sum_{m=1}^{M} \tilde{\ell}(\theta, \tilde{x}_m; \tilde{y}_m), \tag{28}$$

where

$$\tilde{\ell}(\theta, \tilde{x}_m; \tilde{y}_m) := \log p^\eta(\tilde{y}_m - f_\theta(\tilde{x}_m)) + \log p^x_\theta(\tilde{x}_m) \quad \forall m = 1, \dots, M. \tag{29}$$

Because p^η is Gaussian, $(\tilde{x}, \theta) \mapsto \tilde{\ell}(\tilde{x}, \theta; \tilde{y})$'s Hessian $\nabla^2_{(\theta,\tilde{x})}\tilde{\ell}$ does not depend on $\tilde{y}$. Consequently, if $\tilde{\ell}$ is $\tilde{\lambda}$-strongly concave and its gradient is $\tilde{L}$-Lispchitz,

$$\tilde{\lambda} I \preceq -\nabla^2_{(\theta,\tilde{x})}\tilde{\ell} \preceq \tilde{L} I \tag{30}$$

for some positive $\tilde{\lambda}$ and $\tilde{L}$, then ℓ is $(M\tilde{\lambda})$-strongly concave and has a $(M\tilde{L})$-Lipschitz gradient. Combining this fact with Corollary 10, we obtain the following:

Corollary 11 *Let the conditions in Theorem 9's premise hold. Suppose that $(\theta_0, q_0) = (0, \delta_0)$, ℓ is as in (28) with $\tilde{l}$ twice-differentiable and satisfying (30), and $\epsilon > 0$. If*

$$h \le \frac{C_1 \epsilon^2}{\tilde{d}_x M}, \qquad N \ge \frac{C_2 \tilde{d}_x}{\epsilon^2}, \qquad K \ge \frac{C_3 \tilde{d}_x}{\epsilon^2} \log\left(\frac{1}{\tilde{d}_x^{1/2}\epsilon}\right),$$

for some constants $C_1, C_2, C_3 > 0$ independent of M, then

$$\mathsf{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\theta_*, Q_*^N)) \le C\epsilon,$$

for another constant $C > 0$ independent of M.

The corresponding cost is $\mathcal{O}(\epsilon^{-4} \log(\epsilon^{-1})[\text{eval. cost of } \tilde{\ell}])$, where we are ignoring any dependence on $\tilde{d}_x$ which is usually small for these models. In particular, the cost depends on M only through the evaluation cost of ℓ, which cannot be avoided. We can, however, mitigate this by parellizing/distributing over both particles and datapoints, so lowering the running time to $\mathcal{O}(\epsilon^{-2} \log(\epsilon^{-1})[\text{eval. cost of } \tilde{\ell}])$.

The corollary also provides some theoretical support for the idea that the PGD algorithm and its derivatives can work well for problems in which d_x is large, even when N is relatively modest, provided that the structure of the model allows the log-likelihood to be decomposed as a sum of functions of low-dimensional contributions arising from individual datapoints. This appeared to be the case in several examples explored in Kuntz et al. (2023, Section 3), including a 'generator network' model for which $d_x \approx 2.5 \times 10^6$ but good performance was obtained with only 10 particles (similar results were also obtained in Lim et al. (2024, Section 6.3) using their version of PGD). The model, however, lies beyond the scope of Corollary 11 because, in that case, f_θ in (29) is a non-convex neural network and θ encompasses its weights. Moreover, Kuntz et al. (2023) adapted PGD's step sizes to account for this and also subsampled ℓ's gradient to facilitate efficient implementation on a single GPU. Regardless, we view the corollary as a tentative first step in explaining the behaviour observed in Kuntz et al. (2023).

3.3 PGD's Derivation and Theorem 9's Proof

We obtain PGD by approximating the gradient flow (5) via the McKean–Vlasov SDE (6). The first step is to approximate the law, q_t, of X_t in (6) with the empirical distribution $Q_t^N := N^{-1} \sum_{n=1}^N \delta_{X_t^n}$ of N i.i.d. copies $X_t^1, \ldots, X_t^N$ of X_t. To obtain these copies, consider

$$\mathrm{d}\theta_t = \frac{1}{N} \sum_{n=1}^N \int \nabla_\theta \ell(\theta_t, x) q_t^n(\mathrm{d}x)\, \mathrm{d}t, \quad \mathrm{d}X_t^n = \nabla_x \ell(\theta_t, X_t^n)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t^n, \quad \forall n \in [N], \quad (31)$$

where $[N] := \{1, \ldots, N\}$, $W^1, \ldots, W^N$ denote N independent Brownian motions, $q_t^n :=$ Law(X_t^n) for each n, and we assume that the initial condition X_0^n of each particle $(X_t^n)_{t \geq 0}$ is drawn independently from the law q_0 of $(X_t)_{t \geq 0}$'s initial condition X_0. Because the particles all satisfy the same differential equation and have the same initial distribution, they share the same law (in particular, q_t^n is independent of n). Hence, for any given n, we can re-write the leftmost equation in (31) as $\mathrm{d}\theta_t = \int \nabla_\theta \ell(\theta_t, x) q_t^n(\mathrm{d}x)\, \mathrm{d}t$. Putting this equation together with that for X_t^n in (31), we recover (6) with X_t^n replacing X_t therein. In other words, if $(\theta_t, X_t^1, \ldots, X_t^N)_{t \geq 0}$ solves (31), then $(\theta_t, X_t^n)_{t \geq 0}$ solves (6) for any n. Consequently, Proposition 8 tells us that, for any n, $(\theta_t, q_t^n)_{t \geq 0}$ is a classical solution to the gradient flow (5). Exploiting this link and Theorem 7, we can bound the distance between (θ_t, Q_t^N) and (θ_*, Q_*^N), where $Q_*^N := N^{-1} \sum_{n=1}^N \delta_{X_*^n}$ and $X_*^1, \ldots, X_*^N$ denotes N i.i.d. particles with law π_{θ_*}:

Lemma 12 *Under Assumptions 1(ii–iii) and 3–4, if q_0 belongs to $\mathcal{P}_2(\mathbb{R}^{d_x})$, then (31) has a strong solution $(\theta_t, X_t^1, \ldots, X_t^N)_{t \geq 0}$ and the solution satisfies*

$$\mathsf{d}((\theta_t, Q_t^N), (\theta_*, Q_*^N)) \leq \mathsf{d}((\theta_0, q_0), (\theta_*, \pi_{\theta_*})) e^{-\lambda t} \quad \forall t \geq 0.$$

Proof. The argument for the existence of solutions for (31) is similar to that for (6) in Proposition 8's proof, and we skip it. The error bound follows by optimally coupling the two sets of particles; see Appendix B.2 for details. ∎

The next step entails approximating $N^{-1} \sum_{n=1}^N q_t^n(\mathrm{d}x)$ in (31) with the particle's empirical distribution Q_t^N; which leads us to the following SDE:

$$\mathrm{d}\Theta_t^N = \frac{1}{N} \sum_{n=1}^N \nabla_\theta \ell(\Theta_t^N, \bar{X}_t^n)\, \mathrm{d}t, \quad \mathrm{d}\bar{X}_t^n = \nabla_x \ell(\Theta_t^N, X_t^n)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t^n, \qquad \forall n \in [N]. \quad (32)$$

where $\bar{X}_0^n = X_0^n$ for each n in $[N]$. We can bound the approximation error between (31) and (32) as follows:

Lemma 13 (Bound on the space discretization error) *Under Assumptions 3–4, if q_0 belongs to $\mathcal{P}_2(\mathbb{R}^{d_x})$, then (32) has a strong solution $(\Theta_t^N, \bar{X}_t^1, \ldots, \bar{X}_t^N)_{t \geq 0}$ and the solution satisfies*

$$\mathsf{d}((\Theta_t^N, \bar{Q}_t^N), (\theta_t, Q_t^N)) \leq \frac{L\sqrt{2}}{\lambda \sqrt{N}} \sqrt{\|\theta_0\|^2 + \mathbb{E}\left[\|X_0\|^2\right] + \frac{d_x}{\lambda}} \ \ \forall t \geq 0, \ with \ \bar{Q}_t^N := \frac{1}{N} \sum_{n=1}^N \delta_{\bar{X}_t^n}.$$

Proof. (32) is a standard SDE with a Lipschitz drift and diffusion coefficients, so the existence of its solutions is routine (e.g. Øksendal 2013, Theorem 5.2.1). The error bound follows from a synchronous coupling argument; see Appendix B.3 for details. ∎

Lastly, we obtain PGD by discretizing (32) in time using the Euler–Maruyama scheme with step size h. Given Lemmas 12 and 13, Theorem 9 follows from the triangle inequality for d and the following bound on the discretization error:

Lemma 14 (Bound on the time discretization error) *Under Assumptions 3–4. If q_0 belongs to $\mathcal{P}_2(\mathbb{R}^{d_x})$, $h \leq 1/(\lambda + L)$, and $A_{0,h}$ is as in Theorem 9, then:*

$$\mathsf{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\Theta_{Kh}^N, \bar{Q}_{Kh}^N)) \leq \sqrt{h} A_{0,h}.$$

Proof. The bound follows from a synchronous coupling argument; see Appendix B.4. ∎

Other, more refined, discretization procedures such as Randomized Midpoint Discretization (Shen and Lee, 2019; He et al., 2020) might result in an algorithm with better dependencies of the error on the dimension d_x.

4. Discussion

In this paper, we theoretically validated PGD (Algorithm 1) by bounding its error for models with strongly concave log-likelihoods. In order to do this we analyzed the theoretical properties of both PGD and its continuous-time infinite-particle limit: the gradient flow (5). To study the latter, we extended certain inequalities well-known in the optimal transport and optimization literatures and several results interlinking them. In doing so, we obtain conditions weaker than strong concavity of the likelihood under which the flow converges at an exponential rate to the set $\mathcal{M}_*$ of pairs $(\theta_*, \pi_{\theta_*})$ of marginal likelihood maximizers θ_* and matching posterior distributions π_{θ_*}. Under the additional assumptions that the model log-likelihood is strongly concave and its gradient is Lipschitz continuous, we then proved explicit non-asymptotic error bounds for PGD.

Aside from evidencing PGD's well-foundedness, our bounds enable theoretical comparisons between PGD and other maximum marginal likelihood algorithms. For instance, it is interesting to compare the behaviour of the EM algorithm itself and its many variants with PGD. Caprio and Johansen (2024) derived non-asymptotic error bounds for the EM and the gradient EM algorithms under the xLSI. Comparing their EM error bound (Corollary 13 therein) to PGD's (Theorem 9) suggests that the latter is slower, even in the limit of infinite particles. This is consistent with the interpretation of EM as coordinate descent applied to free energy and PGD as a gradient descent: on Euclidean spaces, coordinate descent is typically faster, in terms of number of iterations, in the special cases where it can be implemented (Beck and Tetruashvili, 2013). Balakrishnan et al. (2017) and Kunstner et al. (2021) also provide non-asymptotic analyses of EM, but a comparison with their results is not immediate.

Akyildiz et al. (2025) analyze the IPLA algorithm which can be viewed as a modification of PGD more amenable to standard analysis. They obtain very similar scalings in ϵ, d_x, d_θ for h, N, and K necessary to achieve an error of ϵ under warm-start conditions (see Section 3.6

therein) to those obtained here. The main difference between these bounds is that N must be proportional to $d_x\epsilon^{-2}$ in our case and proportional to $d_\theta\epsilon^{-2}$ in the case of IPLA. Their dependence on d_θ stems from the extra noise term featuring IPLA's parameter updates; see Algorithm 1 and the proof of Proposition 3 in Akyildiz et al. (2025). We conjecture that the dependence on d_x present in our results but not those of Akyildiz et al. (2025) simply reflects the fact that our bounds controlling the error of both the parameter estimates and the particle approximation, while theirs only control the former.

Similarly, our bounds could help settle the question of whether it is possible to 'accelerate' PGD and achieve faster convergence rates via the use of momentum (Lim et al., 2024) as its the case for both gradient descent (Nesterov, 1983) and the unadjusted Langevin algorithm (Ma et al., 2021). In particular, Lim et al. (2024, Theorem 4.1) shows that, under the xLSI, the free energy F along the limiting flow of such a 'momentum-enriched' version of PGD converges exponentially fast to F's infimum. Given Theorem 4, applying our extension of the Talagrand inequality, we can immediately prove convergence in d-distance for flow studied in Lim et al. (2024). By following arguments similar to those in Section 3.3, it might be possible to then also obtain similar results for the momentum-enriched PGD and compare the algorithms' performances.

While our motivation for this work was understanding the theoretical properties of PGD, we find the results in Section 2 interesting in their own right and we believe they might find applications beyond the study of algorithms for maximum marginal likelihood estimation like PGD. For instance, the xLSI is used in Lim and Johansen (2024) to theoretically motivate a novel algorithm for semi-implicit variational inference. Additionally, for models with λ-strongly concave log-likelihoods, Theorem 6 gives us a (to the best of our knowledge) novel upper bound on the optimal marginal log-likelihood:

$$\log(Z_*) \leq \frac{I(\theta, q)}{2\lambda} - F(\theta, q) \quad \forall (\theta, q) \in \mathcal{M}_2^1. \tag{33}$$

Similarly, for models satisfying the xLSI, Theorem 4 yields a bound on the d-distance between (θ, q) and the optimal set $\mathcal{M}_*$:

$$\lambda \mathsf{d}((\theta, q), \mathcal{M}_*) \leq \sqrt{I(\theta, q)} \quad \forall (\theta, q) \in \mathcal{M}_2^1. \tag{34}$$

It would be interesting to explore the extent to which the right-hand side of (33), or of (34) prove informative for models of practical interest. Moreover, for such models, (34) hints at potential (also to the best of our knowledge) new variational inference methods that would minimize $(\theta, \phi) \mapsto I(\theta, q_\phi)$, where $(q_\phi)_{\phi \in \Phi}$ denotes a parametrized family, rather than the conventional approach of minimizing $(\theta, \phi) \mapsto F(\theta, q_\phi)$.

Acknowledgments

Data sharing is not applicable to this article as no new data were generated or analyzed.

Appendix A. Proofs for Section 2

A.1 de Bruijn's Identity

(17) extends de Bruijn's identity (cf. Bakry et al. (2014, Proposition 5.2.2)). Assumption 2 is sufficient for it to hold:

Lemma 15 *If Assumption 2 holds, then equation* (17) *also holds.*

Proof. The regularity of $(\theta_t, q_t)_{t \geq 0}$ allow us to exchange differentiation and integration by the Leibniz rule and compute

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \log(q_t(x)) q_t(\mathrm{d}x) = \int (\log(q_t(x)) + 1) \nabla_x \cdot \left[q_t(x) \nabla_x \log \left(\frac{q_t(x)}{\rho_{\theta_t}(x)} \right) \right] \mathrm{d}x$$

$$= - \int \left\langle \nabla_x \log(q_t(x)), \nabla_x \log \left(\frac{q_t(x)}{\rho_{\theta_t}(x)} \right) \right\rangle q_t(\mathrm{d}x)$$

where we integrated by parts, and

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \ell(\theta_t, x) q_t(\mathrm{d}x) = \int \left\langle \nabla_\theta \ell(\theta_t, x), \dot{\theta}_t \right\rangle q_t(\mathrm{d}x) + \int \ell(\theta_t, x) \nabla_x \cdot \left[q_t(x) \nabla_x \log \left(\frac{q_t(x)}{\rho_{\theta_t}(x)} \right) \right] \mathrm{d}x$$

$$= \left\| \int \nabla_\theta \ell(\theta_t, x) q_t(\mathrm{d}x) \right\|^2 - \int \left\langle \nabla_x \log(\rho_{\theta_t}(x)), \nabla_x \log \left(\frac{q_t(x)}{\rho_{\theta_t}(x)} \right) \right\rangle q_t(\mathrm{d}x).$$

Upon re-arranging, this shows $\frac{\mathrm{d}}{\mathrm{d}t} F(\theta_t, q_t) = -I(\theta_t, q_t)$. ∎

A.2 Proof of Theorem 4

Here some arguments are similar to those in Otto and Villani (2000). Throughout this appendix, we assume that Assumptions 1–2 hold. Let (θ, q) be any point in $\mathcal{M}_2$. As we show in Lemma 16 below,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{d}((\theta_t, q_t), (\theta, q)) \leq \sqrt{I(\theta_t, q_t)} \quad \forall t > 0. \tag{35}$$

By the theorem's premise, the measures $(\rho_\theta(\mathrm{d}x))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the xLSI or, in other words,

$$\sqrt{I(\theta, q)} \leq \frac{I(\theta, q)}{\sqrt{2\lambda[F(\theta, q) - F_*]}} \quad \forall (\theta, q) \in \mathcal{M}_2^1.$$

By the de Bruijn's identity (17), we find that

$$\frac{I(\theta_t, q_t)}{2\sqrt{F(\theta_t, q_t) - F_*}} = -\frac{\mathrm{d}}{\mathrm{d}t} \sqrt{F(\theta_t, q_t) - F_*} \quad \forall t > 0.$$

Combining the above three (in)equalities, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{d}((\theta_t, q_t), (\theta, q)) \leq \sqrt{I(\theta_t, q_t)} \leq \frac{I(\theta_t, q_t)}{\sqrt{2\lambda[F(\theta_t, q_t) - F_*]}} = -\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\frac{2[F(\theta_t, q_t) - F_*]}{\lambda}} \quad \forall t > 0.$$

Integrating over t in (t, t') for $t' \geq t \geq 0$, yields

$$\mathsf{d}((\theta_{t'}, q_{t'}), (\theta, q)) - \mathsf{d}((\theta_t, q_t), (\theta, q)) \leq \sqrt{\frac{2[F(\theta_t, q_t) - F_*]}{\lambda}} - \sqrt{\frac{2[F(\theta_{t'}, q_{t'}) - F_*]}{\lambda}}.$$

If we set $(\theta, q) := (\theta_t, q_t)$, then

$$\mathsf{d}((\theta_t, q_t), (\theta_{t'}, q_{t'})) \leq \sqrt{\frac{2[F(\theta_t, q_t) - F_*]}{\lambda}} - \sqrt{\frac{2[F(\theta_{t'}, q_{t'}) - F_*]}{\lambda}} \quad \forall t' \geq t \geq 0. \qquad (36)$$

As we show in Lemma 17 below, there exists an increasing sequence $t_1 < t_2 < \dots$ approaching ∞ such that

$$(\theta_{t_n}, q_{t_n}) \to (\theta_*, q_*) \quad \text{as} \quad n \to \infty, \qquad (37)$$

for some (θ_*, q_*) belonging to the optimal set $\mathcal{M}_*$ in the topology induced by d on $\mathcal{M}_2$. Because $F(\theta_{t_n}, q_{t_n})$ converges to F_* as $n \to \infty$ (Theorem 2) and

$$(\theta', q') \mapsto \mathsf{d}((\theta, q), (\theta', q'))$$

is a continuous function on $(\mathcal{M}_2, \mathsf{d})$ for any given (θ, q) in $\mathcal{M}_2$, setting $(t, t') := (0, t_n)$ in (36) and taking the limit $n \to \infty$, then implies that

$$\mathsf{d}((\theta_0, q_0), \mathcal{M}_*) \leq \sqrt{\frac{2[F(\theta_0, q_0) - F_*]}{\lambda}};$$

since $(\theta_0, q_0) \in \mathcal{M}_2$ is arbitrary, the claim follows.

Lemma 16 (35) *holds.*

Proof. Let $(\theta, q) \in \mathcal{M}_2$, take $(\theta_t, q_t)_{t \geq 0}$ as per Assumption 2 and define the velocity field $v_t(x) := \nabla_x \log(q_t(x)/\rho_{\theta_t}(x))$. Ambrosio et al. (2008, Theorem 8.4.7) shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{d}_{W_2}(q_t, q)^2 = 2\int \langle v_t(x_1), x_1 - x_2 \rangle \, \mathrm{d}\varrho(x_1, x_2) \qquad (38)$$

where ϱ it an optimal transport plan for (q_t, q). By an application of the Cauchy–Schwarz inequality and the definition of the Wasserstein-2 distance we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{d}_{W_2}(q_t, q)^2 \leq 2\sqrt{\int \|x_1 - x_2\|^2 \, \mathrm{d}\varrho(x_1, x_2) \int \|v_t(x)\|^2 \, q_t(\mathrm{d}x)}$$

$$= 2\mathsf{d}_{W_2}(q_t, q)\sqrt{\int \|v_t(x)\|^2 \, q_t(\mathrm{d}x)}$$

18

(this inequality may also be proved directly using the Benamou–Brenier's formula). On the other hand, $\frac{\mathrm{d}}{\mathrm{d}t}\|\theta_t - \theta\|^2 = 2\left\langle \dot{\theta}_t, \theta_t - \theta\right\rangle \leq 2\|\dot{\theta}_t\|\|\theta_t - \theta\|$, and by these estimates, and then again by Cauchy–Schwarz,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{d}((\theta_t, q_t), (\theta, q)) = \frac{\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{d}((\theta_t, q_t), (\theta, q))^2}{2\mathsf{d}((\theta_t, q_t), (\theta, q))} \leq \sqrt{\|\dot{\theta}_t\|^2 + \int \left\|v_t(x)\right\|^2 q_t(\mathrm{d}x)} = \sqrt{I(\theta_t, q_t)}.$$

∎

Lemma 17 *For any increasing sequence $t_1 < t_2 < \ldots$ approaching ∞, (37) holds for some (θ_*, q_*) in $\mathcal{M}_*$.*

Proof. Inequality (36) tells us that the sequence is Cauchy in $(\mathcal{M}_2^1, \mathsf{d})$ and, consequently, in $(\mathcal{M}_2 := \mathbb{R}^{d_\theta} \times \mathcal{P}_2(\mathbb{R}^{d_x}), \mathsf{d})$. Because both $(\mathbb{R}^{d_\theta}, \mathsf{d}_E)$ and $(\mathcal{P}_2(\mathbb{R}^{d_x}), \mathsf{d}_{W_2})$ are complete metric spaces (Villani, 2009, Theorem 6.18), so is $(\mathcal{M}_2, \mathsf{d})$. Hence, $(\theta_{t_n}, q_{t_n})_{n \in \mathbb{N}}$ converges to a limit $(\theta_\infty, q_\infty)$ in $\mathcal{M}_2$. As we show in Lemma 18 below, the free energy F is lower semicontinuous on $(\mathcal{M}_2, \mathsf{d})$. Consequently, Theorem 2 implies that

$$F(\theta_\infty, q_\infty) \leq \liminf_{n \to \infty} F(\theta_{t_n}, q_{t_n}) = F_*,$$

where the equality holds because $(\theta_{t_n}, q_{t_n})_{n \in \mathbb{N}} \subseteq \mathcal{M}_2^1$. Since it also holds $F(\theta_\infty, q_\infty) \geq F_*$, the claim follows. ∎

Lemma 18 *F is lower semicontinuous on $(\mathcal{M}_2, \mathsf{d})$.*

Proof. Given any π in $\mathcal{P}_2(\mathbb{R}^{d_x})$, the duality formula for the KL divergence (e.g., Ambrosio et al., 2008, Lemma 9.4.4) reads

$$\mathrm{KL}(q\|\pi) = \sup\left\{\int \phi(x)q(\mathrm{d}x) - \int (e^{\phi(x)} - 1)\pi(\mathrm{d}x) : \phi \in \mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})\right\},$$

where $\mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})$ denotes the set of real valued continuous bounded functions with domain $\mathbb{R}^{d_x}$. Consequently,

$$F(\theta, q) = \mathrm{KL}(q\|\pi_\theta) - \log(Z_\theta) = \sup\{G(\theta, q, \phi) : \phi \in \mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})\} - \log(Z_\theta),$$

where $G(\theta, q, \phi) := \int \phi(x)q(\mathrm{d}x) - \int (e^{\phi(x)} - 1)\pi_\theta(\mathrm{d}x)$. Because Assumption 1(i) implies that $\theta \mapsto \log(Z_\theta)$ is continuous and the pointwise supremum of a family of lower semicontinuous (l.s.c.) functions is l.s.c., we need only show that $(\theta, q) \mapsto G(\theta, q, \phi)$ is l.s.c. for each ϕ in $\mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})$. Since the Wasserstein-2 topology is finer than the weak topology, it suffices to show that $(\theta, q) \mapsto G(\theta, q, \phi)$ is continuous in the weak topology for any given ϕ in $\mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})$. Consider a sequence $(\theta_n, q_n)_{n \in \mathbb{N}}$ such that $\theta_n \to \theta$ and $q_n \to q$ weakly. By definition of weak convergence, $\int \phi(x)q_n(\mathrm{d}x) \to \int \phi(x)q(\mathrm{d}x)$. Furthermore, since $x \mapsto (e^{\phi(x)} - 1) \in \mathcal{C}_b(\mathbb{R}^{d_x}, \mathbb{R})$, we have $|\int (e^{\phi(x)} - 1)\pi_{\theta_n}(\mathrm{d}x) - \int (e^{\phi(x)} - 1)\pi_\theta(\mathrm{d}x)| \leq c_\phi \int |\pi_{\theta_n}(x) - \pi_\theta(x)|\,\mathrm{d}x$ for some $c_\phi < \infty$. Since $\theta \mapsto \pi_\theta$ is continuous by Assumption 1(i), an application of Scheffé's lemma concludes the proof. ∎

A.3 Proof of Theorem 6

We preface the proof of Theorem 6 with three auxiliary results which describe variations of the free energy along geodesics in $(\mathcal{M}_2, \mathsf{d})$.

Definition 19 *A curve* $\gamma(t) : t \in [0,1] \mapsto \mathcal{M}_2$ *is a (constant speed) geodesic if*

$$\mathsf{d}(\gamma(s), \gamma(t)) = (t - s)\mathsf{d}(\gamma(0), \gamma(1)), \quad \forall 0 \le s \le t \le 1.$$

For a probability measure $\mu \in \mathcal{P}(\mathbb{R}^{d_x})$ and a measurable map T with domain $\mathbb{R}^{d_x}$, $T_{\#}\mu = \mu \circ T^{-1}$ indicates the pushforward of μ by T.

Lemma 20 *A curve* $\gamma(t) : t \in [0,1] \mapsto \mathcal{M}_2$ *is a geodesic if and only if* $\gamma(t) = (\gamma_\theta(t), \gamma_q(t))$, *where* γ_θ *is a geodesic in* $(\mathbb{R}^{d_\theta}, \|\cdot\|)$ *and* γ_q *is a geodesic in* $(\mathcal{P}_2(\mathbb{R}^{d_x}), \mathsf{d}_{W_2})$. *In particular, if* $\gamma(t)$ *is a geodesic in* $(\mathcal{M}_2, \mathsf{d})$ *connecting* (θ, q) *and* (θ', q') *then* $\gamma(t) = (\gamma_\theta(t) = (1 - t)\theta + t\theta', \gamma_q(t) = (h_t)_{\#}\varrho)$ *where* ϱ *is a Wasserstein* $- 2$ *optimal transport plan for* q, q' *and* $h_t(x, x') = (1 - t)x + tx'$. *Furthermore, if* q *has a density w.r.t. the Lebesgue measure, we can also write* $\gamma_q(t) = ((1 - t)id + t\nabla_x\Phi)_{\#}q$ *for some convex function* Φ.

Proof. The first claim is almost immediate from the definition, also see Burago et al. (2001, Lemma 3.6.4). The second claim follows from the first using the characterization of geodesics in $(\mathcal{P}_2(\mathbb{R}^{d_x}), \mathsf{d}_{W_2})$—see e.g. Santambrogio (2015, Theorem 5.27). The last claim is Brenier's Theorem (Villani, 2009, Theorem 9.4). ∎

The following lemma uses similar arguments to Cheng and Bartlett (2018, Lemma 9), and it connects the strong concavity of the log-likelihood function with strong geodesic convexity of F.

Lemma 21 *If Assumption 3 holds, F is λ-strongly geodesically convex i.e. for every* $(\theta, q), (\theta', q')$ *there is constant speed geodesic* $\gamma : [0,1] \to \mathcal{M}_2$ *connecting* $\gamma(0) = (\theta, q)$ *and* $\gamma(1) = (\theta', q')$ *such that* $t \mapsto F(\gamma(t))$ *is λ-strongly convex:*

$$F(\gamma(t)) \le (1 - t)F(\theta, q) + tF(\theta', q') - \frac{\lambda t(1 - t)}{2}\mathsf{d}((\theta, q), (\theta', q'))^2. \tag{39}$$

Proof. Let $\gamma = (\gamma_\theta, \gamma_q)$ be a geodesic in $\mathcal{M}_2$ connecting (θ, q) and (θ', q'), and recall that by Lemma 20 γ_θ and γ_q are geodesics in $(\mathbb{R}^{d_\theta}, \|\cdot\|)$ and $(\mathcal{P}_2(\mathbb{R}^{d_x}), \mathsf{d}_{W_2})$, respectively. Since $F(\theta, q) = \int \log(q(x))q(\mathrm{d}x) - \int \ell(\theta, x)q(\mathrm{d}x)$, and since $q \mapsto \int \log(q(x))q(\mathrm{d}x)$ is convex along γ_q by Santambrogio (2015, Theorem 7.28), we just need to show that $V(\theta, q) := -\int \ell(\theta, x)q(\mathrm{d}x)$ is λ-strongly convex along γ. Using the representation in Lemma 20 above

and then applying Assumption 3 we obtain

$$
\begin{aligned}
V(\gamma(t)) &= -\int \ell(\gamma_\theta(t), x)\gamma_q(t)(x)\,\mathrm{d}x \\
&= -\int \ell((1-t)\theta + t\theta', (1-t)x + tx')\varrho(x, x')\,\mathrm{d}x\,\mathrm{d}x' \\
&\leq -\int\int ((1-t)\ell(\theta, x) + t\ell(\theta', x'))\varrho(x, x')\,\mathrm{d}x\,\mathrm{d}x' \\
&\quad - \frac{\lambda t(1-t)}{2}\int \left(\|\theta - \theta'\|^2 + \|x - x'\|^2\right)\varrho(x, x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= (1-t)V(\theta, q) + tV(\theta', q') - \frac{\lambda t(1-t)}{2}\mathsf{d}((\theta, q), (\theta', q'))^2.
\end{aligned}
$$

∎

The next result supplies an estimate for the right lower derivative of the free energy along geodesics in $\mathcal{M}_2$ (to compare with Villani, 2009, Theorem 20.1, or Otto and Villani, 2000, Equation 51).

Lemma 22 *Let γ be a geodesic connecting $(\theta, q), (\theta', q') \in \mathcal{M}_2^1$. If ℓ is differentiable, then*

$$
\liminf_{t \to 0^+} \frac{F(\gamma(t)) - F(\gamma(0))}{t} \geq \left\langle \nabla_x \Phi - \mathrm{id}, \nabla_x \delta_q F(\theta, q) \right\rangle_q + \left\langle \theta' - \theta, \nabla_\theta F(\theta, q) \right\rangle
$$

where $\delta_q F(\theta, q) = \log(q(x)) - \ell(\theta, x)$ is F's first variation and $\nabla_x \Phi$ is the optimal transport map from q to q'.

Proof. Let us consider the geodesic $\gamma(t)$ given by $\gamma(t) := (\theta_t, q_t)$ with $q_t(\mathrm{d}x) := ((1-t)x + t\nabla_x \Phi(x))\#q(\mathrm{d}x)$ and $\theta_t = (1-t)\theta + t\theta'$ — see Lemma 20. Since Φ is a convex function (Lemma 20), by Alexandrov's theorem (Villani, 2009, Theorem 14.24) Φ has a second derivative almost everywhere. Let ∇_x^2 denote the Hessian with respect to the variable x. We use the change of variables $x \to (1-t)x + t\nabla_x \Phi(x)$ and then the facts $\theta_t = (1-t)\theta + t\theta'$ and $q_t(\mathrm{d}x) := ((1-t)x + t\nabla_x \Phi(x))\#q(\mathrm{d}x) \Rightarrow q_t((1-t)x + t\nabla_x \Phi(x)) = q(x)/\det((1-t)I_{d_x} + t\nabla_x^2 \Phi(x))$, so

$$
\begin{aligned}
F(\gamma(t)) &= F(\theta_t, q_t) = \int q_t(x)(\log(q_t(x)) - \ell(\theta_t, x))\,\mathrm{d}x \\
&= \int q_t((1-t)x + t\nabla_x \Phi(x)) \log(q_t((1-t)x + t\nabla_x \Phi(x)) - q_t((1-t)x + t\nabla_x \Phi(x)) \\
&\quad \cdot \ell(\theta_t, (1-t)x + t\nabla_x \Phi(x)) \det((1-t)I_{d_x} + t\nabla_x^2 \Phi(x))\,\mathrm{d}x \\
&= \int q(x) \log\left(\frac{q(x)}{\det((1-t)I_{d_x} + t\nabla_x^2 \Phi(x))}\right)\,\mathrm{d}x - q(x)\ell(\theta_t, (1-t)x + t\nabla_x \Phi(x))\,\mathrm{d}x \\
&=: \int q(x) \log(q(x))\,\mathrm{d}x + F_2(\gamma(t)) + F_3(\gamma(t)).
\end{aligned}
$$

Now, by Fatou's lemma,

$$
\liminf_{t \to 0^+} \frac{F_2(\theta_t, q_t) - F_2(\theta_0, q_0)}{t} \geq \int \liminf_{t \to 0^+} \frac{\log(\det(I_{d_x})) - \log(\det((1-t)I_{d_x} + t\nabla_x^2 \Phi(x)))}{t} q(\mathrm{d}x)
$$

21

and by positive semi-definiteness of $(1 - t)I_{d_x} + t\nabla_x^2\Phi(x)$, the arithmetic mean–geometric mean inequality yields that

$$\det((1 - t)I_{d_x} + t\nabla_x^2\Phi(x)) \leq \left(\frac{\mathrm{Tr}((1 - t)I_{d_x} + t\nabla_x^2\Phi(x))}{d_x}\right)^{d_x} = \left((1 - t) + \frac{t}{d_x}\Delta_x\Phi(x)\right)^{d_x}$$

where Δ_x denotes the Laplacian. It follows that

$$\liminf_{t\to 0^+} \frac{F_2(\theta_t, q_t) - F_2(\theta_0, q_0)}{t} \geq \int \liminf_{t\to 0^+} \left[-\frac{d_x}{t}\log\left((1 - t) + \frac{t}{d_x}\Delta_x\Phi(x)\right)\right]q(\mathrm{d}x)$$

$$= -\int(\Delta_x\Phi(x) - d_x)q(\mathrm{d}x).$$

Next, we estimate $F_3(\gamma(t))$'s lower derivative as

$$\liminf_{t\to 0^+} \frac{F_3(\theta_t, q_t) - F_3(\theta_0, q_0)}{t} \geq \int \liminf_{t\to 0^+} \left[\frac{\ell(\theta, x) - \ell(\theta_t, (1 - t)x + t\nabla_x\Phi(x))}{t}\right]q(\mathrm{d}x)$$

$$= -\int \langle \nabla\ell(\theta, x), (\theta' - \theta, \nabla_x\Phi(x) - x)\rangle q(\mathrm{d}x)$$

where the limit exchange is again justified by Fatou's lemma. Putting these results together, integrating by parts, and using the divergence theorem then gives

$$\liminf_{t\to 0^+} \frac{F(\theta_t, q_t) - F(\theta_0, q_0)}{t}$$

$$\geq -\int q(x)(\Delta_x\Phi(x) - d_x)\,\mathrm{d}x - \int q(x)\langle\nabla\ell(\theta, x), (\theta' - \theta, \nabla_x\Phi(x) - x)\rangle\,\mathrm{d}x$$

$$= \int \langle\nabla_x q(x), \nabla_x\Phi(x) - x\rangle\,\mathrm{d}x - \int q(x)\langle\nabla\ell(\theta, x), (\theta' - \theta, \nabla_x\Phi(x) - x)\rangle\,\mathrm{d}x$$

$$= \int \langle\nabla_x q(x), \nabla_x\Phi(x) - x\rangle - q(x)\langle\nabla_x\ell(\theta, x), \nabla_x\Phi(x) - x\rangle\,\mathrm{d}x$$

$$- \int \langle\theta' - \theta, \nabla_\theta\ell(\theta, x)\rangle q(x)\,\mathrm{d}x$$

$$= \int \left\langle\nabla_x\log\left(\frac{q(x)}{\rho_\theta(x)}\right), \nabla_x\Phi(x) - x\right\rangle q(x)\,\mathrm{d}x - \int \langle\theta' - \theta, \nabla_\theta\ell(\theta, x)\rangle q(x)\,\mathrm{d}x$$

∎

Proof of Theorem 6. Take any two points (θ, q) and (θ', q') in $\mathcal{M}_2^1$ and consider a geodesic $\gamma = (\gamma_\theta, \gamma_q)$ connecting those. Since under Assumption 1, ℓ is differentiable, taking the right lower derivative w.r.t. t at 0 in (39) and using Lemma 22 we obtain

$$\langle\nabla_x\Phi - id, \nabla_x\delta_q F(\theta, q)\rangle_q + \langle\theta' - \theta, \nabla_\theta F(\theta, q)\rangle$$

$$\leq \liminf_{t\to 0^+} \frac{F(\gamma(t)) - F(\gamma(0))}{t} \leq F(\theta', q') - F(\theta, q) - \frac{\lambda}{2}\mathrm{d}((\theta, q), (\theta', q'))^2. \tag{40}$$

Because $\|\nabla_x \Phi - id\|_q^2 = \mathsf{d}_{W_2}(q, q')^2$, setting $(\theta', q') := (\theta_*, \pi_{\theta_*})$ and using the Cauchy–Schwartz inequality yields

$$
\begin{aligned}
F(\theta, q) - F_* &\leq - \left\langle \nabla_x \Phi - id, \nabla_x \delta_q F(\theta, q) \right\rangle_q - \left\langle \theta' - \theta, \nabla_\theta F(\theta, q) \right\rangle - \frac{\lambda}{2} \mathsf{d}((\theta, q), (\theta', q'))^2 \\
&\leq + \mathsf{d}((\theta, q), (\theta', q')) \sqrt{\left\| \nabla_x \delta_q F(\theta, q) \right\|_q^2 + \left\| \nabla_\theta F(\theta, q) \right\|^2} - \frac{\lambda}{2} \mathsf{d}((\theta, q), (\theta', q'))^2 \\
&= + \mathsf{d}((\theta, q), (\theta', q')) \sqrt{I(\theta, q)} - \frac{\lambda}{2} \mathsf{d}((\theta, q), (\theta', q'))^2.
\end{aligned}
$$

Now we use Young's inequality $ab \leq a^2/2\epsilon + \epsilon b^2/2$ with $\epsilon = \lambda > 0$ on $\mathsf{d}((\theta, q), (\theta', q')) \sqrt{I(\theta, q)}$ and the claim follows. ∎

A.4 Proof of Theorem 7

Let $t > 0$. Consider a geodesic $\gamma = (\gamma_\theta, \gamma_q)$ connecting the points (θ_t, q_t) and $(\theta_*, \pi_{\theta_*})$ in $\mathcal{M}_2^1$. Setting $(\theta, q) = (\theta_t, q_t)$ and $(\theta', q') = (\theta_*, \pi_{\theta_*})$ in (40), and since under Assumptions 1(iii) and 3 the xLSI holds (Theorem 6), we can use the extended Talagrand-type inequality (Theorem 4) to obtain

$$
\begin{aligned}
&\left\langle \nabla_x \Phi - id, \nabla_x \delta_q F(\theta_t, q_t) \right\rangle_{q_t} + \left\langle \theta_\star - \theta_t, \nabla_\theta F(\theta_t, q_t) \right\rangle \\
&\leq F_* - F(\theta_t, q_t) - \frac{\lambda}{2} \mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*}))^2 \leq -\lambda \mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*}))^2.
\end{aligned}
$$

Let ϱ be the optimal transport plan for (q_t, π_{θ_*}) and let $v_t(x) = \nabla_x \log(q_t(x)/\rho_{\theta_t}(x))$. Recall that $\varrho = (id \times \nabla_x \Phi)_\# q_t$ by Brenier's Theorem. Combining the above inequality with (38) we write

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*}))^2 &= 2 \int \left\langle v_t(x_1), x_1 - x_2 \right\rangle \mathrm{d}\varrho(x_1, x_2) + 2 \left\langle \dot{\theta}_t, \theta_t - \theta_* \right\rangle \\
&= 2 \left\langle \nabla_x \Phi - id, \nabla_x \delta_q F(\theta_t, q_t) \right\rangle_{q_t} + 2 \left\langle \theta_\star - \theta_t, \nabla_\theta F(\theta_t, q_t) \right\rangle \leq -2\lambda \mathsf{d}((\theta_t, q_t), (\theta_*, \pi_{\theta_*}))^2,
\end{aligned}
$$

upon which the result follows via Grönwall's inequality. ∎

Appendix B. Proofs for Section 3

B.1 Proof of Proposition 8

We break down the proof of Proposition 8 into three steps: (i) in Lemma 23 we prove that under our assumptions the SDE (6) has an unique strong solution; (ii) in Lemma 24 we prove some regularity properties for the law of the solution; and finally (iii) in Lemma 25 we prove that such law provides a classical solution to (5) satisfying Assumption 2. In this section, Assumptions 1(ii) and 4 are taken to hold.

Lemma 23 *For all $T > 0$, if (θ, X) belongs to $\mathbb{R}^{d_\theta} \times \mathcal{L}^2(\mathbb{R}^{d_x})$, the SDE (6) has an unique strong solution $(\theta_t, X_t)_{t \leq T}$ on $[0, T]$ on such that $(\theta_0, X_0) = (\theta, X)$.*

Proof. This is an adaptation of the arguments in Carmona (2016, Theorem 1.7), Chaintron and Diez (2022, Proposition 1), and Lim et al. (2024, Proposition 3.1). Fix any $T > 0$, θ_0 in $\mathbb{R}^{d_\theta}$, and X_0 in $\mathcal{L}^2(\mathbb{R}^{d_x})$. Let $\mathcal{C}([0,T], \mathcal{S})$ be the space of continuous functions from $[0, T]$ to a metric space $\mathcal{S}$. It is straightforward to check that our Lipschitz assumption on ℓ's gradient, implies that $\ell(\theta, x)$ is ν_t-integrable for any t, θ, and ν in $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$. Consequently, for any given such ν,

$$\mathrm{d}(\theta_t^\nu, X_t^\nu) = b^\nu(\theta_t^\nu, X_t^\nu, t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad (\theta_0^\nu, X_0^\nu) = (\theta_0, X_0), \tag{41}$$

with $b^\nu(\theta, x, t) := [\int \nabla_\theta \ell(\theta, z') \nu_t(\mathrm{d}z), \nabla_x \ell(\theta, x)]^\intercal$ and $\sigma := \sqrt{2}\mathrm{diag}(0_{d_\theta}, 1_{d_x})$, is well-defined. Assumption 4 and Jensen's inequality imply that b^ν is Lipschitz in (θ, x), uniformly over t:

$$\left\|b^\nu(\theta, x, t) - b^\nu(\theta', x', t)\right\|^2 \leq \int \left\|\nabla_\theta \ell(\theta, z) - \nabla_\theta \ell(\theta', z)\right\| \nu_t(\mathrm{d}z)^2 + \left\|\nabla_x \ell(\theta, x) - \nabla_x \ell(\theta', x')\right\|^2$$

$$\leq L^2 \left\|\theta - \theta'\right\|^2 + L^2\left\|(\theta, x) - (\theta', x')\right\|^2$$

$$\leq 2L^2\left\|(\theta, x) - (\theta', x')\right\|^2, \quad (\theta, x), (\theta', x') \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}, t \leq T.$$

Moreover, for any point (θ', x') in $\mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}$,

$$\left\|b^\nu(\theta, x, t)\right\|^2$$

$$\leq \int \left\|\nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta', x')\right\|^2 \nu_t(\mathrm{d}x) + \left\|\nabla_x \ell(\theta, x) - \nabla_x \ell(\theta', x')\right\|^2 + \left\|\nabla \ell(\theta', x')\right\|^2$$

$$\leq 2L\left\|(\theta, x) - (\theta', x')\right\|^2 + a_T, \quad (\theta, x) \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}, t \leq T,$$

where $a_T := L\sup_{t \leq T}\left\|\theta - \theta'\right\|^2 + \int \left\|x - x'\right\|^2 \nu_t(\mathrm{d}x) + \left\|\nabla \ell(\theta', x')\right\|^2 < \infty$. Consequently, Øksendal (2013, Theorem 5.2.1.) tells us that (41) has a unique strong solution $(\theta_t^\nu, X_t^\nu)_{t \leq T}$ over $[0, T]$. Furthermore, $(\mathrm{Law}(X_t^\nu))_{t \leq T}$ belongs to $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$: combining the above with Jensen's inequality,

$$\mathbb{E}\left[\|\theta_t^\nu\|^2 + \|X_t^\nu\|^2\right] = \mathbb{E}\left[\left\|\int_0^t b^\nu(\theta_s^\nu, X_s^\nu, s)\,\mathrm{d}s\right\|^2\right]$$

$$\leq 2T\left(L\int_0^t \mathbb{E}\left[\left\|\theta_s^\nu - \theta'\right\|^2 + \left\|X_s^\nu - x'\right\|^2\right]\mathrm{d}s + a_T\right)$$

$$\leq 2T\left(2L\int_0^T \mathbb{E}\left[\|\theta_s^\nu\|^2 + \|X_s^\nu\|^2\right]\mathrm{d}s + a_T + TL(\|\theta'\|^2 + \|x'\|^2)\right), \quad \forall t \leq T;$$

and it follows from Grönwall's inequality that $\mathbb{E}\left[\|\theta_t^\nu\|^2 + \|X_t^\nu\|^2\right] \leq 2T(a_T + TL(\|\theta'\|^2 + \|x'\|^2)) + 4Le^T T < \infty$ for all $t \leq T$. Now consider the function

$$\Psi_T : \mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x})) \mapsto \mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x})),$$

mapping ν to $(\mathrm{Law}(X_t^\nu))_{t \leq T}$. Note that, if $(\theta_t, X_t)_{t \leq T}$ is a strong solution of (6), then $(\mathrm{Law}(X_t))_{t \leq T}$ is a fixed point of Ψ_T by the latter's definition. Consequently, (6) has no more solutions than Ψ_T has fixed points. On the other hand, if $(q_t)_{t \leq T}$ is a fixed point of

Ψ_T, then there is some $(\theta_t, X_t)_{t \leq T}$ solving (41) with ν equal to q such that $(\mathrm{Law}(X_t))_{t \leq T}$ also equals q. In other words, there is some $(\theta_t, X_t)_{t \leq T}$ that solves (6). In short, to prove the existence and uniqueness of (6)'s solutions, we need only argue that Ψ_T has a unique fixed point. To this end, consider the following metric on $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$:

$$\mathsf{d}_{W_2}^T(\nu, \nu') := \sup_{t \leq T} \mathsf{d}_{W_2}(\nu_t, \nu'_t).$$

If we show that, for some integer k, the k-fold composition Ψ_T^k of Ψ_T with itself is a contraction w.r.t. $\mathsf{d}_{W_2}^T$, the Banach–Caccioppoli fixed-point theorem (Kreyszig, 1978, Theorem 5.1) will imply that Ψ_T has a unique fixed point in $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$. Fix any two elements ν and ν' of $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$, denote with $(\theta_t^\nu, X_t^\nu)_{t \leq T}$ and $(\theta_t^{\nu'}, X_t^{\nu'})_{t \leq T}$ the corresponding SDE (41) solutions, and let c denotes a rolling constant independent of those, which value might change from line to line. Applying in order Jensen's inequality, $\nabla\ell$'s Lipschitz continuity, and the Kantorovich–Rubinstein duality formula,

$$\mathbb{E}\left[\left\|\theta_t^\nu - \theta_t^{\nu'}\right\|^2 + \left\|X_t^\nu - X_t^{\nu'}\right\|^2\right] \leq \mathbb{E}\left[\left\|\int_0^t (b^\nu(\theta_s^\nu, X_s^\nu, s) - b^{\nu'}(\theta_s^{\nu'}, X_s^{\nu'}, s))\,\mathrm{d}s\right\|^2\right]$$

$$\leq T \int_0^t \mathbb{E}\left[\left\|b^\nu(\theta_s^\nu, X_s^\nu, s) - b^{\nu'}(\theta_s^{\nu'}, X_s^{\nu'}, s)\right\|^2 \mathrm{d}s\right]$$

$$\leq cT \int_0^t \mathbb{E}\left[\left\|\theta_s^\nu - \theta_s^{\nu'}\right\|^2 + \left\|X_s^\nu - X_s^{\nu'}\right\|^2 + \mathsf{d}_{W_1}(\nu_s, \nu'_s)^2\right] \mathrm{d}s$$

Because $\mathsf{d}_{W_1} \leq \mathsf{d}_{W_2}$ (Villani, 2009, Remark 6.6),

$$\int_0^t \mathsf{d}_{W_1}(\nu_s, \nu'_s)^2 \,\mathrm{d}s \leq \int_0^t \sup_{r \leq s} \mathsf{d}_{W_1}(\nu_r, \nu'_r)^2 \,\mathrm{d}s \leq \int_0^T \mathsf{d}_s(\nu_{[0,s]}, \nu'_{[0,s]})^2 \,\mathrm{d}s,$$

where $\nu_{[0,s]}, \nu'_{[0,s]}$ respectively denote ν, ν''s restriction to $[0,s]$. Combining the above two and applying Grönwall's inequality, we find that

$$\mathbb{E}\left[\left\|\theta_t^\nu - \theta_t^{\nu'}\right\|^2 + \left\|X_t^\nu - X_t^{\nu'}\right\|^2\right] \leq c_T \int_0^T \mathsf{d}_s(\nu_{[0,s]}, \nu'_{[0,s]})^2 \,\mathrm{d}s$$

with $c_T := cTe^T$. Taking supremums over $t \leq T$, we find that

$$\mathsf{d}_{W_2}^T(\Psi_T(\nu), \Psi_T(\nu'))^2 = \sup_{t \leq T} \mathsf{d}_{W_2}(\nu_t, \nu'_t) \leq \sup_{t \leq T} \mathbb{E}\left[\left\|X_t^\nu - X_t^{\nu'}\right\|^2\right]$$

$$\leq \sup_{t \leq T}\left(\mathbb{E}\left[\left\|\theta_t^\nu - \theta_t^{\nu'}\right\|^2 + \left\|X_t^\nu - X_t^{\nu'}\right\|^2\right]\right) \leq c_T \int_0^T \mathsf{d}_{W_2}^T(\nu_{[0,s]}, \nu'_{[0,s]})^2 \,\mathrm{d}s.$$

Let Ψ_T^k denote the k-fold composition of Ψ_T with itself. Iterating the inequality above k-times yields

$$\mathsf{d}_{W_2}^T(\Psi_T^k(\nu), \Psi_T^k(\nu'))^2 \leq (c_T)^k \int_0^T \frac{s^k}{(k-1)!} \mathsf{d}_{W_2}^T(\nu_{[0,s]}, \nu'_{[0,s]})^2 \,\mathrm{d}s$$

$$\leq \frac{(c_T)^k}{k!} \mathsf{d}_{W_2}^T(\nu, \nu')^2 \,\mathrm{d}s.$$

In particular, for large enough k, Ψ_T^k is a contraction. Because $(\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x})), \mathsf{d}_{W_2}^T)$ is a complete metric space (Sutherland, 2009, Proposition 17.15), the Banach–Caccioppoli fixed-point theorem (Kreyszig, 1978, Theorem 5.1) then implies that Ψ_T has a unique fixed point in $\mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^{d_x}))$. ∎

In the lemma below, for some $T > 0$ we set $\mathbb{R}_T^{d_x} := [0,T) \times \mathbb{R}^{d_x}$ and we write $\mathcal{L}_{\text{loc}}^\infty(\mathbb{R}_T^{d_x})$ for the set of real locally bounded functions on $\mathbb{R}_T^{d_x}$. We let $\mathcal{H}^{j,k}(\mathbb{R}_T^{d_x})$ denote the space of real functions on $\mathbb{R}_T^{d_x}$ for which all components of the weak derivatives $\nabla_t^m \nabla_x^n q$ exist and belong to $\mathcal{L}^\infty(\mathbb{R}_T^{d_x})$ for all $m \leq j, n \leq k$. Here, ∇^i denotes the i-th fold outer product of ∇ with itself. We write that such functions are in $\mathcal{H}_{\text{loc}}^{j,k}(\mathbb{R}_T^{d_x})$ if those weak derivatives only belong to $\mathcal{L}_{\text{loc}}^\infty(\mathbb{R}_T^{d_x})$. In the proof below, we follow an argument in Fan et al. (2023, Lemma C.4), see also Jordan et al. (1998, Theorem 5.1), Mei et al. (2018, Lemma 10.7).

Lemma 24 Law(X_t) *has a Lebesgue density in* $\mathcal{C}^{1,2}(\mathbb{R}_T^{d_x}, \mathbb{R}^+)$ *and* $\theta_t \in \mathcal{C}^1([0,T), \mathbb{R}^{d_\theta})$.

Proof. Set $q_t := \text{Law}(X_t)$. Since $(t,x) \mapsto \nabla_x \log(\rho_{\theta_t}(x))$ is locally integrable in $\mathbb{R}_T^{d_x}$ due Assumption 1 and the fact that $t \mapsto \theta_t$ is continuous, Bogachev et al. (2015, Corollary 6.4.3) shows that $(q_t)_{t \leq T}$ admits a continuous Lebesgue density, further satisfying

$$q_t \in \mathcal{H}_{\text{loc}}^{0,1}(\mathbb{R}_T^{d_x}). \tag{42}$$

We now exploit this first regularity estimate and improve on it with a bootstrap argument. Let φ_{σ^2} denote the Gaussian density with mean 0 and variance σ^2. Let $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^{d_x}, \mathbb{R})$. For fixed $t > 0$ and $y \in \mathbb{R}^{d_x}$, Itô's lemma on $(s, X_s) \mapsto \phi(X_s)\varphi_{\sigma^2 + t-s}(y - X_s)$ yields

$$\int \phi(x)\varphi_{\sigma^2}(y-x) q_t(\mathrm{d}x) = \int_0^t \int (\partial_s \varphi_{\sigma^2+t-s}(y-x)\phi(x) + \langle \nabla_x \ell(\theta_s, x), \nabla_x \phi \rangle \varphi_{\sigma^2+t-s}(y-x)$$
$$- \langle \nabla_x \ell(\theta_s, x), \nabla_x \varphi_{\sigma^2+t-s}(y-x) \rangle \phi(x) + \Delta_x(\phi(x)\varphi_{\sigma^2+t-s}(y-x))) q_s(\mathrm{d}x)\,\mathrm{d}s.$$

Using the heat equation $\partial_s \varphi_{\sigma^2+t-s}(y-x) = -\Delta_x \varphi_{\sigma^2+t-s}(y-x)$ and integrating by parts,

$$\int \phi(x)\varphi_{\sigma^2}(y-x) q_t(\mathrm{d}x) = \int_0^t \int (\langle \nabla_x \ell(\theta_s, x), \nabla_x \phi(x) \rangle + \Delta_x \phi(x)) q_s(x)\varphi_{\sigma^2+t-s}(y-x)$$
$$+ \langle (-\nabla_x \ell(\theta_s, x)\phi(x) + 2\nabla_x \phi(x)) q_s(x), \nabla_x \varphi_{\sigma^2+t-s}(y-x) \rangle \,\mathrm{d}x\,\mathrm{d}s.$$

Let $\sigma^2 \to 0$. Using the weak differentiability of q_t ensured by (42), we deduce

$$\phi(y) q_t(y) = \int_0^t \xi_{1,s} * \varphi_{t-s}(y) + \xi_{2,s} * \varphi_{t-s}(y)\,\mathrm{d}s \quad \forall y \in \mathbb{R}^{d_x}, \tag{43}$$

where $*$ denotes the convolution operator, and we defined $\xi_{1,s}(x) := (\langle \nabla_x \ell(\theta_s, x), \nabla_x \phi(x) \rangle - \Delta_x \phi(x)) q_s(x)$ and $\xi_{2,s}(x) := \nabla_x \cdot (q_s(x)(\nabla_x \ell(\theta_s, x)\phi(x) - 2\nabla_x \phi(x)))$. The next key ingredient is the following implication in Ladyzhenskaia et al. (1968, Chapter 4, (3.1)):

$$\xi_t \in \mathcal{H}^{j,k}(\mathbb{R}_T^{d_x}) \text{ for } 2j + k \leq 2m \Rightarrow \int_0^t \xi_s * \varphi_{t-s}(y)\,\mathrm{d}s \in \mathcal{H}^{j,k}(\mathbb{R}_T^{d_x}) \text{ for } 2j + k \leq 2m + 2 \tag{44}$$

Since $q_t \in \mathcal{H}^{0,1}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T)$, $\phi \in \mathcal{C}^\infty_c(\mathbb{R}^{d_x}, \mathbb{R})$ and the elements of $(t, x) \mapsto \nabla_x \ell(\theta_t, x)$ and $(t, x) \mapsto \nabla_x^2 \ell(\theta_t, x)$ belong to $\mathcal{L}^\infty_{\mathrm{loc}}(\mathbb{R}^{d_x}_T)$ by Assumption 1(ii), the equations (44) and (43) show the chain of implications $q_t \in \mathcal{H}^{0,1}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T) \Rightarrow (\phi q_t) \in \mathcal{H}^{0,2}(\mathbb{R}^{d_x}_T) \Rightarrow q_t \in \mathcal{H}^{0,2}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T) \Rightarrow (\phi q_t) \in \mathcal{H}^{0,3}(\mathbb{R}^{d_x}_T) \Rightarrow q_t \in \mathcal{H}^{0,3}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T)$. Next, since we now proved $q_t \in \mathcal{H}^{0,3}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T)$, by analogous arguments (44) also yields $(\phi q_t) \in \mathcal{H}^{1,3}(\mathbb{R}^{d_x}_T) \Rightarrow q_t \in \mathcal{H}^{1,3}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T) \Rightarrow (\phi q_t) \in \mathcal{H}^{2,3}(\mathbb{R}^{d_x}_T) \Rightarrow q_t \in \mathcal{H}^{2,3}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T)$. Finally, by the Sobolev embedding Theorem (Adams and Fournier, 2003, Theorem 4.12), $q_t \in \mathcal{H}^{2,3}_{\mathrm{loc}}(\mathbb{R}^{d_x}_T) \Rightarrow q_t \in \mathcal{C}^{1,2}(\mathbb{R}^{d_x}_T)$. Since ℓ is continuously differentiable by Assumption 4 and $q_t \in \mathcal{C}^{1,2}(\mathbb{R}^{d_x}_T)$, we also have $\theta_t \in \mathcal{C}^1([0, T), \mathbb{R}^{d_\theta})$, and the claim follows. ∎

Following the, by now standard, argument of Sznitman (1991), as T is arbitrary and for any $T' < T$ the projection of the solution on $[0, T]$ onto $[0, T']$ coincides with the solution obtained by working directly on $[0, T']$, there exists an unique extension on $[0, \infty)$. The regularity properties of the solutions given in Lemma 24 extends to $[0, \infty)$. In fact, the marginals of the solution on $[0, \infty)$ have to agree to the solution on $[0, T)$ for any $T > 0$; such regularity is guaranteed by Lemma 24. Hence, if the regularity failed at some $T' > 0$, we could just take $T = 2T'$, producing a contradiction.

We proved that (6) has an unique strong solution on all $[0, \infty)$ with the regularity that Assumption 2 asks a classical solution to (5) to have. The next lemma closes the circle.

Lemma 25 $(\theta_t, \mathrm{Law}(X_t))_{t \geq 0}$ is a classical solution to (5).

Proof. Let ϕ in $\mathcal{C}^\infty_c(\mathbb{R}^{d_x}, \mathbb{R})$. Itô's lemma shows

$$\phi(X_t) = \phi(X_0) + \int_0^t \left(\langle \nabla_x \ell(\theta_s, X_s), \nabla_x \phi(X_s) \rangle + \Delta_x \phi(X_s) \right) \mathrm{d}s + \int_0^t \langle \nabla_x \phi(X_s), \mathrm{d}W_s \rangle.$$

Consider the system

$$\dot{\theta}_t = \int \nabla_\theta \ell(\theta_t, x) q_t(\mathrm{d}x), \quad M_t := \phi(X_t) - \int_0^t \left(\langle \nabla_x \ell(\theta_s, X_s), \nabla_x \phi(X_s) \rangle + \Delta_x \phi(X_s) \right) \mathrm{d}s$$

with $q_t = \mathrm{Law}(X_t)$. Comparing with the expression for $\phi(X_t)$ above we notice that M_t is a martingale with respect to the natural filtration generated by $(\theta_t, X_t)_{t \geq 0}$, as it corresponds to the Itô integral of an adapted, square integrable process against Brownian motion. Taking expectations and the time derivative in M_t above shows that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \phi(x) q_t(\mathrm{d}x) = - \int \langle \nabla_x \ell(\theta_t, x), \nabla_x \phi(x) \rangle q_t(\mathrm{d}x) + \int \Delta_x \phi(x) q_t(\mathrm{d}x) \quad \forall \phi \in \mathcal{C}^\infty_c(\mathbb{R}^{d_x}, \mathbb{R}).$$

In particular, this shows that $(\theta_t, \mathrm{Law}(X_t))_{t \geq 0}$ is a weak solution to (5). Thanks to the regularity provided by Lemma 24, we can integrate by parts the above to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \phi(x) q_t(x) \, \mathrm{d}x = \int \phi(x)(\Delta_x q_t(x) - \langle \nabla_x \ell(\theta_t, x), \nabla_x q_t(x) \rangle) \, \mathrm{d}x;$$

for all $\phi \in \mathcal{C}^\infty_c(\mathbb{R}^{d_x}, \mathbb{R})$. For each $x \in \mathbb{R}^{d_x}$, we can consider a sequence of bump functions in $\mathcal{C}^\infty_c(\mathbb{R}^{d_x}, \mathbb{R})$ concentrating in the point, showing that (5) holds pointwise and that $(\theta_t, \mathrm{Law}(X_t))_{t \geq 0}$ is a classical solution. ∎

B.2 Proof of Lemma 12

Recall that the particles $(X_t^1, \ldots, X_t^N)$ are i.i.d. with distribution q_t, and that the particles $(X_*^1, \ldots, X_*^N)$ are i.i.d. with distribution π_{θ_*}. Consider the coupling of these random vectors such that each X_*^i is optimally coupled (in the sense of the Wasserstein-2 distance) with X_t^i. In this case, we write

$$
\mathsf{d}((\theta_t, Q_t^N)), (\theta_*, Q_*^N))^2 \leq \|\theta_t - \theta_*\|^2 + \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\|X_t^n - X_*^n\|^2\right]
$$

$$
= \|\theta_t - \theta_*\|^2 + \mathbb{E}\left[\left\|X_t^1 - X_*^1\right\|^2\right] = \mathsf{d}((\theta_t, q_t)), (\theta_*, \pi_{\theta_*}))^2.
$$

The result follows via Theorem 7 and Proposition 8. ∎

B.3 Proof of Lemma 13

Proposition 26 *Assume Assumptions 3–4 and let $(\theta_\dagger, x_\dagger)$ denote ℓ's unique maximizer. Then, we have the following uniform-in-time second moment bound*

$$
\|\theta_t\|^2 + \mathbb{E}\left[\|X_t\|^2\right] \leq 2\left(\|\theta_\dagger\|^2 + \|x_\dagger\|^2 + \|\theta_0 - \theta_\dagger\|^2 + \mathbb{E}\left[\|X_0 - x_\dagger\|^2\right] + \frac{2d_x}{\lambda}\right) \quad \forall t \geq 0.
$$

If $(\theta_\dagger, x_\dagger) = (0,0)$, the above also holds without the factor of 2 in the right hand side.

Proof. Because,

$$
f(z) := \|z\|^2 \Rightarrow \nabla f(z) = 2z, \quad \nabla^2 f(z) = 2I,
$$

by setting

$$
\xi_t := \|X_t - x_\dagger\|^2 + \|\theta_t - \theta_\dagger\|^2 = \|(\theta_t, X_t) - (\theta_\dagger, x_\dagger)\|^2,
$$

applying Itô's formula (Øksendal, 2013, Theorem 4.2.1) and noting that X_t has law q_t, we find

$$
\mathsf{d}\xi_t = 2\left[\left\langle (\theta_t - \theta_\dagger, X_t - x_\dagger), (\mathbb{E}\left[\nabla_\theta \ell(\theta_t, X_t)\right], \nabla_x \ell(\theta_t, X_t))\right\rangle + d_x\right] \mathsf{d}t + 2^{3/2} \langle X_t - x_\dagger, \mathsf{d}W_t\rangle.
$$
(45)

However,

$$
\left\langle (\theta_t - \theta_\dagger, X_t - x_\dagger), (\mathbb{E}\left[\nabla_\theta \ell(\theta_t, X_t)\right], \nabla_x \ell(\theta_t, X_t))\right\rangle \tag{46}
$$

$$
= \left\langle \theta_t - \theta_\dagger, \mathbb{E}\left[\nabla_\theta \ell(\theta_t, X_t)\right]\right\rangle + \left\langle X_t - x_\dagger, \nabla_x \ell(\theta_t, X_t)\right\rangle
$$

$$
= \mathbb{E}\left[\left\langle \theta_t - \theta_\dagger, \nabla_\theta \ell(\theta_t, X_t)\right\rangle\right] + \left\langle X_t - x_\dagger, \nabla_x \ell(\theta_t, X_t)\right\rangle,
$$

and, because ℓ is strongly log-concave by Assumption 3 and $(\theta_\dagger, x_\dagger)$ maximizes it,

$$
\left\langle \nabla \ell(\theta_t, X_t), (\theta_t, X_t) - (\theta_\dagger, x_\dagger)\right\rangle \leq \ell(\theta_t, X_t) - \ell(\theta_\dagger, x_\dagger) - \lambda\xi_t/2 \leq -\lambda\xi_t/2.
$$

(e.g. see pp. 63-64 in Nesterov, 2003). Taking expectations of (46) and applying the above,

$$\mathbb{E}\left[\left\langle (\theta_t - \theta_\dagger, X_t - x_\dagger), (\mathbb{E}\left[\nabla_\theta \ell(\theta_t, X_t)\right], \nabla_x \ell(\theta_t, X_t))\right\rangle\right]$$
$$= \mathbb{E}\left[\left\langle (\theta_t - \theta_\dagger, X_t - x_\dagger), \nabla \ell(\theta_t, X_t)\right\rangle\right] \le -\lambda \mathbb{E}\left[\xi_t\right]/2.$$

In turn, taking expectations of (45) yields $d\mathbb{E}\left[\xi_t\right] \le \left[-\lambda\mathbb{E}\left[\xi_t\right] + 2d_x\right] dt \ \forall t \ge 0$ and

$$\mathbb{E}\left[\xi_t\right] \le e^{-\lambda t}\left(\mathbb{E}\left[\xi_0\right] - \frac{2d_x}{\lambda}\right) + \frac{2d_x}{\lambda} \le \max\left(\mathbb{E}\left[\xi_0\right], \frac{2d_x}{\lambda}\right) \le \mathbb{E}\left[\xi_0\right] + \frac{2d_x}{\lambda} \quad \forall t \ge 0,$$

where we used $e^{-\lambda t} \le 1$. Now the C_p inequality, $\|\theta_t\|^2 \le 2(\|\theta_\dagger\|^2 + \|\theta_t - \theta_\dagger\|^2)$ and similarly for $\mathbb{E}\left[\|X_t\|^2\right]$, gives the desired bound. ∎

Proof of Lemma 13. To prove the desired bound, we start with Itô's formula to write

$$d\left\|\Theta_t^N - \theta_t\right\|^2 = 2\left\langle \Theta_t^N - \theta_t, \frac{1}{N}\sum_{n=1}^N \nabla_\theta \ell(\Theta_t^N, \bar{X}_t^n) - \int \nabla_\theta \ell(\theta_t, x)q_t(dx)\right\rangle dt,$$

$$d\left\|\bar{X}_t^n - X_t^n\right\|^2 = 2\left\langle \bar{X}_t^n - X_t^n, \nabla_x \ell(\Theta_t^N, \bar{X}_t^n) - \nabla_x \ell(\theta_t, X_t^n)\right\rangle dt.$$

Setting $\bar{\xi}_t^N := N^{-1}\sum_{n=1}^N \left\|\bar{X}_t^n - X_t^n\right\|^2 + \left\|\Theta_t^N - \theta_t\right\|^2$, averaging the second equation over n, and adding to the first, we find that

$$d\bar{\xi}_t^N = 2\left\langle \Theta_t^N - \theta_t, \frac{1}{N}\sum_{n=1}^N \nabla_\theta \ell(\Theta_t^N, \bar{X}_t^n) - \int \nabla_\theta \ell(\theta_t, x)q_t(dx)\right\rangle dt$$

$$+ \frac{2}{N}\sum_{n=1}^N \left\langle \bar{X}_t^n - X_t^n, \nabla_x \ell(\Theta_t^N, \bar{X}_t^n) - \nabla_x \ell(\theta_t, X_t^n)\right\rangle dt.$$

Because $X_t^1, \ldots, X_t^N$ all have law q_t,

$$d\bar{\xi}_t^N = \frac{2}{N}\sum_{n=1}^N \left[\left\langle \Theta_t^N - \theta_t, \nabla_\theta \ell(\Theta_t^N, \bar{X}_t^n) - \nabla_\theta \ell(\theta_t, X_t^n)\right\rangle\right.$$

$$\left. + \left\langle \bar{X}_t^n - X_t^n, \nabla_x \ell(\Theta_t^N, \bar{X}_t^n) - \nabla_x \ell(\theta_t, X_t^n)\right\rangle\right] dt + 2G_t^N dt$$

where

$$G_t^N := \left\langle \Theta_t^N - \theta_t, \frac{1}{N}\sum_{n=1}^N \nabla_\theta \ell(\theta_t, X_t^n) - \int \nabla_\theta \ell(\theta_t, x)q_t(dx)\right\rangle$$

$$= \frac{1}{N}\left\langle \Theta_t^N - \theta_t, \sum_{n=1}^N \left[\nabla_\theta \ell(\theta_t, X_t^n) - \mathbb{E}\left[\nabla_\theta \ell(\theta_t, X_t^n)\right]\right]\right\rangle.$$

It then follows from Assumption 3 that

$$d\bar{\xi}_t^N \le [-2\lambda\bar{\xi}_t^N + 2G_t^N]\,dt \ \ \forall t \ge 0, \quad \Rightarrow \quad \frac{d\mathbb{E}\left[\bar{\xi}_t^N\right]}{dt} \le -2\lambda\mathbb{E}\left[\bar{\xi}_t^N\right] + 2\mathbb{E}\left[G_t^N\right] \ \ \forall t \ge 0. \quad (47)$$

As we show below,

$$\left|\mathbb{E}\left[G_t^N\right]\right| \le L\sqrt{\frac{2\mathbb{E}\left[\bar{\xi}_t^N\right]}{N}\left(B_0 + \frac{d_x}{\lambda}\right)}. \quad (48)$$

Because $\frac{d}{dt}\mathbb{E}\left[\bar{\xi}_t^N\right]^{1/2} = 2^{-1}\mathbb{E}\left[\bar{\xi}_t^N\right]^{-1/2}\frac{d}{dt}\mathbb{E}\left[\bar{\xi}_t^N\right]$, (47,48) imply that

$$\frac{d}{dt}\mathbb{E}\left[\bar{\xi}_t^N\right]^{1/2} \le -\lambda\mathbb{E}\left[\bar{\xi}_t^N\right]^{1/2} + L\sqrt{\frac{2}{N}\left(B_0 + \frac{d_x}{\lambda}\right)}.$$

Applying Grönwall's inequality, we obtain that

$$\mathbb{E}\left[\bar{\xi}_t^N\right]^{1/2} \le e^{-\lambda t}\sqrt{\bar{\xi}_0^N} + \frac{(1-e^{-\lambda t})L}{\lambda}\sqrt{\frac{2}{N}\left(B_0 + \frac{d_x}{\lambda}\right)},$$

and the result follows because $\bar{\xi}_0^N = 0$ by construction. ∎

Proof of (48). Applying the Cauchy–Schwarz inequality, we obtain

$$\left|\mathbb{E}\left[G_t^N\right]\right|^2 \le \frac{1}{N^2}\mathbb{E}\left[\left\|\Theta_t^N - \theta_t\right\|^2\right]\mathbb{E}\left[\left\|\sum_{n=1}^N\left[\nabla_\theta\ell(\theta_t, X_t^n) - \mathbb{E}\left[\nabla_\theta\ell(\theta_t, X_t^n)\right]\right]\right\|^2\right]$$

$$\le \frac{\mathbb{E}\left[\bar{\xi}_t^N\right]}{N^2}\sum_{n=1}^N\mathbb{E}\left[\left\|\nabla_\theta\ell(\theta_t, X_t^n) - \mathbb{E}\left[\nabla_\theta\ell(\theta_t, X_t^n)\right]\right\|^2\right] =: \frac{\mathbb{E}\left[\bar{\xi}_t^N\right]}{N^2}\sum_{n=1}^N c_t^n,$$

where the last inequality follows from the independence of $X_t^1, \ldots, X_t^N$. Let X_t' denote a random variable independent of $(X_t^1, \ldots, X_t^n)$ with law q_t. Jensen's inequality and Lipschitz continuity of $\nabla_\theta\ell$ (Assumption 4) imply

$$c_t^n = \mathbb{E}\left[\left\|\nabla_\theta\ell(\theta_t, X_t^n) - \mathbb{E}\left[\nabla_\theta\ell(\theta_t, X_t^n)\right]\right\|^2\right] \le \mathbb{E}\left[\left\|\nabla_\theta\ell(\theta_t, X_t^n) - \nabla_\theta\ell(\theta_t, X_t')\right\|^2\right]$$

$$\le L^2\mathbb{E}\left[\left\|X_t^n - X_t'\right\|^2\right] \le 2L^2\mathbb{E}\left[\left\|X_t'\right\|^2\right].$$

Combining the above two and Proposition 26, we then obtain (48). ∎

B.4 Proof of Lemma 14

The following proof extends the arguments in Durmus and Moulines (2019, Lemma S2) or Chewi (2024, Chapter 4.1). We introduce the linear interpolation of the Euler–Maruyama

discretization of (32):

$$\tilde{X}_t^n = \tilde{X}_{kh}^n + (t - kh)\nabla_x\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N) + \sqrt{2}(W_t^n - W_{kh}^n), \quad \forall n \in [N];$$

$$\tilde{\Theta}_t^N = \tilde{\Theta}_{kh}^N + (t - kh)\frac{1}{N}\sum_{n=1}^N \nabla_\theta\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N);$$

for all $t \in [kh, (k+1)h)$ and $k \in \mathbb{N}$. Let $\tilde{Q}_t^h := N^{-1}\sum_{i=1}\delta_{\tilde{X}_t^n}$. We notice that $(\tilde{\Theta}_{Kh}^N, \tilde{Q}_{Kh}^h)$ coincides in distribution with $(\Theta_K^{N,h}, Q_K^{N,h})$, hence we just need to derive the bound

$$\mathsf{d}((\tilde{\Theta}_{Kh}^N, \tilde{Q}_{Kh}^h), (\Theta_{Kh}^N, \bar{Q}_{Kh}^N)) \leq \sqrt{h}A_{0,h} \quad \forall K \in \mathbb{N}.$$

For all $n \in [N]$, $k \in \mathbb{N}$ and $h > 0$ we compute directly from the defining equations,

$$\left\|\bar{X}_{(k+1)h}^n - \tilde{X}_{(k+1)h}^n\right\|^2 = \left\|\int_{kh}^{(k+1)h}[\nabla_x\ell(\bar{X}_s^n, \Theta_s^N) - \nabla_x\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N)]\,\mathrm{d}s\right\|^2$$

$$- 2h\left\langle \bar{X}_{kh}^n - \tilde{X}_{kh}^n, \nabla_x\ell(\bar{X}_{kh}^n, \Theta_{kh}^N) - \nabla_x\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N)\right\rangle$$

$$- 2\int_{kh}^{(k+1)h}\left\langle \bar{X}_{kh}^n - \tilde{X}_{kh}^n, \nabla_x\ell(\bar{X}_s^n, \Theta_s^N) - \nabla_x\ell(\bar{X}_{kh}^n, \Theta_{kh}^N)\right\rangle\mathrm{d}s + \left\|\bar{X}_{kh}^n - \tilde{X}_{kh}^n\right\|^2,$$

$$\left\|\Theta_{(k+1)h}^N - \tilde{\Theta}_{(k+1)h}^N\right\|^2 = \frac{1}{N}\sum_{n=1}^N\left\|2\int_{kh}^{(k+1)h}[\nabla_\theta\ell(\bar{X}_s^n, \Theta_s^N) - \nabla_\theta\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N)]\,\mathrm{d}s\right\|^2$$

$$- \frac{2h}{N}\sum_{n=1}^N\left\langle \Theta_{kh}^N - \tilde{\Theta}_{kh}^N, \nabla_\theta\ell(\bar{X}_{kh}^n, \Theta_{kh}^N) - \nabla_\theta\ell(\tilde{X}_{kh}^n, \tilde{\Theta}_{kh}^N)\right\rangle$$

$$- \frac{2}{N}\sum_{n=1}^N\int_{kh}^{(k+1)h}\left\langle \Theta_{kh}^N - \tilde{\Theta}_{kh}^N, [\nabla_\theta\ell(\bar{X}_s^n, \Theta_s^N) - \nabla_\theta\ell(\bar{X}_{kh}^n, \Theta_{kh}^N)]\right\rangle\mathrm{d}s + \left\|\Theta_{kh}^N - \tilde{\Theta}_{kh}^N\right\|^2.$$

Averaging the N equations for $\left\|\bar{X}_{(k+1)h}^n - \tilde{X}_{(k+1)h}^n\right\|^2$, adding the one for $\left\|\Theta_{(k+1)h}^N - \tilde{\Theta}_{(k+1)h}^N\right\|^2$ we obtain, with $\bar{Y}_t^n := (\bar{X}_t^n, \Theta_t^N)$, $\tilde{Y}_t^n := (\tilde{X}_t^n, \tilde{\Theta}_t^N)$ and $\xi_t := N^{-1}\sum_{n=1}^N\left\|\tilde{Y}_t^n - \bar{Y}_t^n\right\|^2$,

$$\xi_{(k+1)h} = \xi_{kh} + \frac{1}{N}\sum_{n=1}^N\left\|\int_{kh}^{(k+1)h}[\nabla\ell(\bar{Y}_s^n) - \nabla\ell(\tilde{Y}_{kh}^n)]\,\mathrm{d}s\right\|^2$$

$$- \frac{2h}{N}\sum_{n=1}^N\left\langle \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n, \nabla\ell(\bar{Y}_{kh}^n) - \nabla\ell(\tilde{Y}_{kh}^n)\right\rangle$$

$$- \frac{2}{N}\sum_{n=1}^N\int_{kh}^{(k+1)h}\left\langle \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n, \nabla\ell(\bar{Y}_s^n) - \nabla\ell(\bar{Y}_{kh}^n)\right\rangle\mathrm{d}s. \tag{49}$$

Now, adding and subtracting $\nabla \ell(\bar{Y}_{kh}^n)$, applying Jensen's inequality, expanding the resulting integrand and further applying Young's inequalities,

$$\left\| \int_{kh}^{(k+1)h} [\nabla \ell(\bar{Y}_s^n) - \nabla \ell(\tilde{Y}_{kh}^n)] \, \mathrm{d}s \right\|^2 \leq 2h^2 \left\| \nabla \ell(\bar{Y}_{kh}^n) - \nabla \ell(\tilde{Y}_{kh}^n) \right\|^2$$

$$+ 2h \int_{kh}^{(k+1)h} \left\| \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\|^2 \mathrm{d}s. \qquad (50)$$

Furthermore, under Assumptions 3 and 4 we have, by Nesterov (2003, Theorem 2.1.12), there holds the co-coercivity property

$$\langle \nabla \ell(y') - \nabla \ell(y), y' - y \rangle \geq \frac{\iota}{2} \| y - y' \|^2 + \frac{\| \nabla \ell(y') - \nabla \ell(y) \|^2}{\lambda + L},$$

where $\iota := 2\lambda L / (\lambda + L)$. It follows that, whenever $h < 1/(\lambda + L)$, we have

$$\frac{1}{N} \sum_{n=1}^{N} \left[2h^2 \left\| \nabla \ell(\bar{Y}_{kh}^n) - \nabla \ell(\tilde{Y}_{kh}^n) \right\|^2 - 2h \left\langle \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n, \nabla \ell(\bar{Y}_{kh}^n) - \nabla \ell(\tilde{Y}_{kh}^n) \right\rangle \right]$$

$$\leq -\frac{h\iota}{N} \sum_{n=1}^{N} \left\| \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n \right\|^2 = -h\iota \xi_{kh} \qquad (51)$$

Combining (49,50,51), we obtain

$$\xi_{(k+1)h} \leq (1 - \iota h) \xi_{kh} - \frac{2}{N} \sum_{n=1}^{N} \left[\int_{kh}^{(k+1)h} \left\langle \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n, \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\rangle \mathrm{d}s \right.$$

$$\left. + 2h \int_{kh}^{(k+1)h} \left\| \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\|^2 \mathrm{d}s \right].$$

To deal with the remaining terms, fix any $\epsilon > 0$. Applying Young's inequality, we obtain

$$\frac{1}{N} \sum_{n=1}^{N} \left| \left\langle \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n, \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\rangle \right|$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\epsilon}{2} \left\| \bar{Y}_{kh}^n - \tilde{Y}_{kh}^n \right\|^2 + \frac{1}{2\epsilon} \left\| \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\|^2 \right]$$

$$= \frac{\epsilon}{2} \xi_{kh} + \frac{1}{2\epsilon N} \sum_{n=1}^{N} \left\| \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\|^2.$$

Putting the above two together, we obtain for any $\epsilon > 0$ and $k \in \mathbb{N}$,

$$\xi_{(k+1)h} \leq (1 - \iota h + h\epsilon/2) \xi_{kh} + \left(2h + \frac{1}{2\epsilon} \right) \int_{kh}^{(k+1)h} \frac{1}{N} \sum_{n=1}^{N} \left\| \nabla \ell(\bar{Y}_s^n) - \nabla \ell(\bar{Y}_{kh}^n) \right\|^2 \mathrm{d}s. \quad (52)$$

We will show below that if $h < 1/(\lambda + L)$ then, for all $s \in [kh, (k+1)h)$,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla\ell(\bar{Y}_s^n) - \nabla\ell(\bar{Y}_{kh}^n)\right\|^2\right] \leq 220hL^2(L^2h(B_0 + d_x/\lambda + d_x)). \tag{53}$$

Hence, taking expectations in (52), choosing $\epsilon = \iota$ using the bound above and the fact that $\xi_0 = 0$ we obtain

$$\mathbb{E}\left[\xi_{(k+1)h}\right] \leq (2h + 2/\iota)220h^2(L^2h(B_0 + d_x/\lambda + d_x)\sum_{j=1}^{k}(1 - \iota h/2)^j$$

$$\leq \frac{4h + 4/\iota}{\iota}220hL^2(L^2h(B_0 + d_x/\lambda + d_x).$$

∎

Proof of (53). Let $\{\mathcal{F}_t; t \geq 0\}$ be the filtration generated by $(\theta_t, \bar{X}_t^n)_{n=1}^N$ and denote $\mathbb{E}_{kh}[\cdot]$ expectation conditional on $\mathcal{F}_{kh}$. We first show that whenever $h < 1/(\lambda + L)$

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{kh}\left[\left\|\nabla\ell(\bar{Y}_s^n) - \nabla\ell(\bar{Y}_{kh}^n)\right\|^2\right] \leq \frac{1}{N} \sum_{n=1}^{N} 220hL^2(L^2h\|\bar{Y}_{kh}^n\|^2 + d_x), \tag{54}$$

after which the result follows from taking expectations and applying Proposition 27. Since

$$\mathbb{E}_{kh}\left[\left\|\bar{X}_s^n - \bar{X}_{kh}^n\right\|^2\right] = \mathbb{E}_{kh}\left[\left\|\int_{kh}^{s} \nabla_x\ell(\Theta_r^N, \bar{X}_r^n)\,\mathrm{d}r + \sqrt{2}(W_s^n - W_{kh}^n)\right\|^2\right]$$

$$\leq 2(s - kh)\int_{kh}^{s} \mathbb{E}_{kh}\left[\left\|\nabla_x\ell(\Theta_r^N, \bar{X}_r^n)\right\|^2\right]\mathrm{d}r + 4d_x(s - kh)$$

for all $n \in [N]$, and

$$\left\|\Theta_s^N - \Theta_{kh}^N\right\|^2 = \left\|\frac{1}{N}\sum_{n=1}^{N}\int_{kh}^{s}\nabla_\theta\ell(\Theta_r^N, \bar{X}_r^n)\,\mathrm{d}r\right\|^2 \leq 2(s - kh)\frac{1}{N}\sum_{n=1}^{N}\int_{kh}^{s}\left\|\nabla_\theta\ell(\Theta_r^N, \bar{X}_r^n)\right\|^2\mathrm{d}r,$$

adding up and since Assumption 4 implies $\left\|\nabla\ell(\bar{Y}_r^n)\right\|^2 \leq L^2\left\|\bar{Y}_r^n\right\|^2$, we obtain

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{kh}\left[\left\|\bar{Y}_s^n - \bar{Y}_{kh}^n\right\|^2\right] \leq 2hL^2\int_{kh}^{s}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{kh}\left[\left\|\bar{Y}_r^n\right\|^2\right]\mathrm{d}r + 4d_xh$$

$$\leq 4hL^2\int_{kh}^{s}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{kh}\left[\left\|\bar{Y}_r^n - \bar{Y}_{kh}^n\right\|^2\right]\mathrm{d}r + 4h^2L^2\mathbb{E}_{kh}\left[\left\|\bar{Y}_{kh}^n\right\|^2\right] + 4d_xh$$

and by Grönwall's lemma

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{kh}\left[\left\|\bar{Y}_s^n - \bar{Y}_{kh}^n\right\|^2\right] \leq \frac{1}{N}\sum_{n=1}^{N}\exp(4L^2h^2)(4L^2h^2\left\|\bar{Y}_{kh}^n\right\|^2 + 4d_xh).$$

If $h < 1/(\lambda + L)$, $\exp(4L^2h^2) \leq 55$, and (54) then follows by Assumption 4. ∎

Proposition 27 *Assume Assumptions 3–4. Then, for all $t \geq 0$,*

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\|X_t^n\|^2 + \left\|\Theta_t^N\right\|^2\right] \leq 2\left(\|\theta_0\|^2 + \mathbb{E}\left[\left\|\bar{X}_0\right\|^2\right] + \frac{d_x}{\lambda}\right)$$

Proof. Recalling that here $(\theta_\dagger, x_\dagger) = (0, 0)$, this argument is very similar to that supporting Proposition 26. For each $n \in [N]$,

$$d\|X_t^n\|^2 = -2\left\langle X_t^n, \nabla_x \ell(\Theta_t^N, X_t^n)\right\rangle dt + \sqrt{2}\, dW_t^n + 2d_x\, dt$$

$$d\left\|\Theta_t^N\right\|^2 = -\frac{2}{N}\sum_{n=1}^{N}\left\langle \Theta_t^N, \nabla_\theta \ell(\Theta_t^N, X_t^n)\right\rangle dt$$

now averaging the N equations for $\|X_t^n\|^2$ and then adding $\left\|\Theta_t^N\right\|^2$, taking expectations and time derivatives, with $\xi_t^N = N^{-1}\sum_{n=1}^{N}\|X_t^n\|^2 + \left\|\Theta_t^N\right\|^2$,

$$\frac{d}{dt}\mathbb{E}\left[\xi_t^N\right] = -2\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left\langle(X_t^n, \Theta_t^N), \nabla\ell(\Theta_t^N, X_t^n)\right\rangle\right] + 2d_x \leq -2\lambda\mathbb{E}\left[\xi_t^N\right] + 2d_x,$$

and Grönwall's inequality provides the conclusion. ∎

References

Robert A. Adams and John JF Fournier. *Sobolev Spaces*. Elsevier, 2nd edition, 2003.

Ö. Deniz Akyildiz, Francesca R. Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis. Interacting particle Langevin algorithm for maximum marginal likelihood estimation. *ESAIM: Probability and Statistics*, 2025. In press.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10:1116–1135, 2000.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*. Lecture Notes in Mathematics Springer, 1985.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 40: 77–120, 2017.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Vladimir I. Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*. American Mathematical Society, 2015.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

Rocco Caprio and Adam M Johansen. Fast convergence of the Expectation Maximization algorithm under a logarithmic Sobolev inequality. eprint 2407.17949, arXiv, 2024.

René Carmona. *Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications*. SIAM, 2016.

Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. I. Models and methods. *Kinetic and Related Models*, 15(6), 2022.

Rujian Chen. *Approximate Bayesian Modeling with Embedded Gaussian Processes*. PhD thesis, Massachusetts Institute of Technology, 2023.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 186–211, 2018.

Sinho Chewi. Log-concave sampling. Book draft, 2024. URL https://chewisinho.github.io.

Francesca R. Crucinio. A mirror descent approach to maximum likelihood estimation in latent variable models. eprint 2501.15896, arXiv, 2025.

Arnak S. Dalalyan. Theoretical Guarantees for Approximate Sampling from Smooth and Log-Concave Densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39:2–38, 1977.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'Institut Henri Poincaré: Probabilités et statistiques*, volume 49, pages 1183–1203, 2013.

Randal Douc, Éric Moulines, and David Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC press, 2014.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27:1551 – 1587, 2017.

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

Paula C. Encinar, Ö. Deniz Akyildiz, and Francesca R. Crucinio. Proximal interacting particle Langevin algorithms. eprint 2406.14292, arXiv, 2024.

Zhou Fan, Leying Guan, Yandi Shen, and Yihong Wu. Gradient flows for empirical Bayes in high-dimensional linear models. eprint 2312.12708, arXiv, 2023.

Nicolas Fournier. Convergence of the empirical measure in expected Wasserstein distance: non-asymptotic explicit bounds in $\mathbb{R}^d$. *ESAIM: Probability and Statistics*, 27:749–775, 2023.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

Alexander Genkin, David D. Lewis, and David Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4): 1061–1083, 1975.

Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Advances in Neural Information Processing Systems*, 33:7366–7376, 2020.

Aapo Hyvärinen. Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22(1):49–67, 1998.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12:307–392, 2019.

Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1978.

Frederik Kunstner, Raunak Kumar, and Mark Schmidt. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 3295–3303. PMLR, 2021.

Juan Kuntz, Jen Ning Lim, and Adam M. Johansen. Particle algorithms for maximum likelihood training of latent variable models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 5134–5180, 2023.

Olga A Ladyzhenskaia, Vsevolod A. Solonnikov, and Nina N. Ural'tseva. *Linear and quasi-linear equations of parabolic type*. American Mathematical Society, 1968.

Jen Ning Lim and Adam M Johansen. Particle semi-implicit variational inference. In *Proceedings of 38th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pages 123954–123990, 2024.

Jen Ning Lim, Juan Kuntz, Samuel Power, and Adam M. Johansen. Momentum particle maximum likelihood. In *Proceedings of 41st International Conference on Machine Learning (ICML)*, volume 235, pages 29816–29871. PMLR, 2024.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *Advances in Neural Information Processing Systems*, 29: 2378–2386, 2016.

Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942 – 1992, 2021.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-Layer Neural Networks. *Proceedings of the National Academy of Sciences*, 115 (33):7665–7671, 2018.

Andriy Mnih and Russ R. Salakhutdinov. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, pages 1257–1264, 2007.

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer Netherlands, 1998.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2013.

Paul F. V. Oliva and Ö. Deniz Akyildiz. Kinetic interacting particle Langevin Monte Carlo. eprint 2407.05790, arXiv, 2024.

Félix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.

Boris T. Polyak. Gradient methods for the minimisation of functionals (in Russian). *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3:643–653, 1963.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3.1, pages 157–164, 1956.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser/Springer, 2015.

Louis Sharrock, Daniel Dodd, and Christopher Nemeth. Tuning-free maximum likelihood training of latent variable models via coin betting. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32:2100–2111, 2019.

Wilson A. Sutherland. *Introduction to Metric and Topological Spaces*. Oxford University Press, 2nd edition, 2009.

Alain-Sol Sznitman. Topics in Propagation of Chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer, Berlin, 1991.

Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis*, 6:587–600, 1996.

Nicolas G. Trillos, Bamdad Hosseini, and Daniel Sanz-Alonso. From optimization to sampling through gradient flows. *Notices of the American Mathematical Society*, 70(6), 2023.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302, 2021.

Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2009.

Tim Y. J. Wang, Juan Kuntz, and Ö. Deniz Akylidìz. Training latent diffusion models with interacting particle algorithms. eprint 2505.12412, arXiv, 2025.

Antoine Wehenkel and Gilles Louppe. Diffusion priors in variational autoencoders. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.