Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions

Umut Şimşekli[1], Antoine Liutkus[2], Szymon Majewski[3], Alain Durmus[4]

1: LTCI, Télécom Paristech, Université Paris-Saclay, 75013, Paris, France

2: Inria and LIRMM, Montpellier, France

3: Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland

4: CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France

Abstract

By building up on the recent theory that established the connection between implicit generative modeling and optimal transport, in this study, we propose a novel parameter-free algorithm for learning the underlying distributions of complicated datasets and sampling from them. The proposed algorithm is based on a functional optimization problem, which aims at finding a measure that is close to the data distribution as much as possible and also expressive enough for generative modeling purposes. We formulate the problem as a gradient flow in the space of probability measures. The connections between gradient flows and stochastic differential equations let us develop a computationally efficient algorithm for solving the optimization problem, where the resulting algorithm resembles the recent dynamics-based Markov Chain Monte Carlo algorithms. We provide formal theoretical analysis where we prove finite-time error guarantees for the proposed algorithm. Our experimental results support our theory and shows that our algorithm is able to capture the structure of challenging distributions.

Contents

1 Introduction

Implicit generative modeling (IGM) [1, 2] has become very popular recently and has proven successful in various fields; variational auto-encoders (VAE) [3] and generative adversarial networks (GAN) [4] being its two well-known examples. The goal in IGM can be briefly described as learning the underlying probability measure of a given dataset, denoted as $\nu \in \mathcal{P}(\Omega)$, where $\mathcal{P}$ is the space of probability measures on the measurable space $(\Omega, \mathcal{A})$, $\Omega \subset \mathbb{R}^d$ is a domain and $\mathcal{A}$ is the associated Borel σ-field.

Given a set of data points $\{y_1, \ldots, y_P\}$ that are assumed to be independent and identically distributed (i.i.d.) samples drawn from ν, the implicit generative framework models the data points as the output of a measurable map, i.e. $y = T(x)$, with $T : \Omega_\mu \mapsto \Omega$. Here, the inputs x are generated from a known and easy to sample source measure μ on Ω_μ (e.g. Gaussian or uniform measures), and the outputs $T(x)$ should match the unknown target measure ν on Ω.

Learning generative networks has witnessed several groundbreaking contributions in the recent years. Motivated by this fact, there has been an interest in illuminating the theoretical foundations of VAEs and GANs [5, 6]. It has been shown that these implicit models have close connections with the theory of Optimal Transport (OT) [7]. As it turns out, OT brings new light on the generative modeling problem: there have been several extensions of VAEs [8, 9] and GANs [10, 11, 12, 13], which exploit the links betwen OT and IGM.

OT studies whether it is possible to transform samples from a source distribution μ to a target distribution ν. From this perspective, an ideal generative model is simply a transport map from μ to ν. This can be written by using some 'push-forward operators': we seek a mapping T that 'pushes μ onto ν', and is formally defined as $\nu(A) = \mu(T^{-1}(A))$ for all Borel sets $A \subset \mathcal{A}$. If this relation holds, we denote the push-forward operator

$T_\#$, such that $T_\#\mu = \nu$. Provided mild conditions on these distributions hold (notably μ is non-atomic [7]), existence of such a transport map is guaranteed; however, it remains a challenge to construct it in practice.

One common point between VAE and GAN is to adopt an approximate strategy and consider transport maps that belong to a parametric family T_ϕ with $\phi \in \Phi$. Then, they aim at finding the best parameter $\phi^\star$ that would give $T_{\phi^\star\#}\mu \approx \nu$. This is typically achieved by attempting to minimize the following optimization problem: $\phi^\star = \arg\min_{\phi\in\Phi} \mathcal{W}_2(T_{\phi\#}\mu, \nu)$, where $\mathcal{W}_2$ denotes the Wasserstein distance that will be properly defined in Section 2. It has been shown that [14] OT-based GANs [10] and VAEs [8] both use this formulation with different parameterizations and different equivalent definitions of $\mathcal{W}_2$; however, their resulting algorithms still lack theoretical understanding.

In this study, we follow a completely different approach for IGM and we seek to estimate a transport map between source μ and target ν that is *nonparametric*, but rather iteratively augmented, always increasing the quality of the fit along iterations. Formally, we take T_t as the constructed transport map at time t, and define $\mu_t = T_t\#\mu$ as the corresponding output distribution. Our objective is to build the maps so that μ_t will converge to the solution of a functional optimization problem, defined through a gradient flow in the Wasserstein space. Informally, we will consider a gradient flow that have the following form:

$$\partial_t \mu_t = -\nabla_{\mathcal{W}_2}\Big\{\mathrm{Cost}(\mu_t, \nu) + \mathrm{Reg}(\mu_t)\Big\}, \qquad \mu_0 = \mu, \tag{1}$$

where the functional Cost computes a discrepancy between μ_t and ν, Reg denotes a regularization functional, and $\nabla_{\mathcal{W}_2}$ denotes a notion of gradient with respect to a probability measure in the $\mathcal{W}_2$ metric for probability measures[1]. If this flow can be simulated, one would hope for $\mu_t = T_t\#\mu$ to converge to the minimum of the functional optimization problem: $\min_\mu(\mathrm{Cost}(\mu, \nu) + \mathrm{Reg}(\mu))$.

We construct a gradient flow where we choose the Cost functional as the *sliced Wasserstein distance* ($\mathcal{SW}_2$) and the Reg functional as the negative entropy. The $\mathcal{SW}_2$ distance is equivalent to the $\mathcal{W}_2$ distance [15] and has important computational implications since it can be expressed as an average of (one-dimensional) projected optimal transportation costs whose analytical expressions are available.

We first show that, with the choice of $\mathcal{SW}_2$ and the negative-entropy functionals as the overall objective, we obtain a valid gradient flow that has a solution path $(\mu_t)_t$, and the probability density functions of this path solve a particular partial differential equation, which has close connections with stochastic differential equations. Even though the gradient flows in Wasserstein spaces cannot be solved in general, by exploiting this connection, we are able to develop a practical algorithm that provides approximate solutions to the gradient flow and is reminiscent of stochastic gradient Markov Chain Monte Carlo (MCMC) methods [16, 17]. We provide finite-time error guarantees for the proposed algorithm and show explicit dependence of the error to the algorithm parameters.

Apart from its nice theoretical properties, the proposed algorithm has also significant practical importance: (i) it has low computationally requirements and can be easily run on an everyday laptop CPU, (ii) it has a strong potential for privacy preserving applications since it only requires random projections of the data,

[1]This gradient flow is similar to the usual Euclidean gradient flows, i.e. $\partial_t x_t = -\nabla(f(x_t) + r(x_t))$, where f is typically the data-dependent cost function and r is a regularization term. The (explicit) Euler discretization of this flow results in the well-known gradient descent algorithm for solving $\min_x(f(x) + r(x))$.

rather than the data itself. Our experiments on both synthetic and real datasets support our theory and illustrate the advantages of the algorithm in challenging scenarios.

2 Preliminaries and Technical Background

2.1 Wasserstein distance, optimal transport maps and Kantorovich potentials

For two probability measures $\mu, \nu \in \mathcal{P}_2(\Omega)$, $\mathcal{P}_2(\Omega) = \{\mu \in \mathcal{P}(\Omega) : \int_\Omega \|x\|^2 \mu(\mathrm{d}x) < +\infty\}$, the 2-Wasserstein distance is defined as follows:

$$\mathcal{W}_2(\mu,\nu) \triangleq \Big\{ \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int_{\Omega\times\Omega} \|x-y\|^2 \gamma(dx,dy) \Big\}^{1/2}, \tag{2}$$

where $\mathcal{C}(\mu,\nu)$ is called the set of *transportation plans* and defined as the set of probability measures γ on $\Omega \times \Omega$ satisfying for all $A \in \mathcal{A}$, $\gamma(A \times \Omega) = \mu(A)$ and $\gamma(\Omega \times A) = \nu(A)$, i.e. the marginals of γ coincide with μ and ν. From now on, we will assume that Ω is a compact subset of $\mathbb{R}^d$.

In the case where Ω is finite, computing the Wasserstein distance between two probability measures turns out to be a linear program with linear constraints, and has therefore a dual formulation. Since Ω is a Polish space (i.e. a complete and separable metric space), this dual formulation can be generalized as follows [7, Theorem 5.10]:

$$\mathcal{W}_2(\mu,\nu) = \sup_{\psi \in \mathrm{L}^1(\mu)} \Big\{ \int_\Omega \psi(x)\mu(dx) + \int_\Omega \psi^c(x)\nu(dx) \Big\}^{1/2}, \tag{3}$$

where ψ^c denotes the c-conjugate of ψ and is defined as follows: $\psi^c(y) \triangleq \{\inf_{x\in\Omega} \|x-y\|^2 - \psi(x)\}$. The functions ψ that realize the supremum in (3) are called the Kantorovich potentials between μ and ν. Provided that μ satisfies a mild condition, we have the following nice uniqueness result.

Theorem 1 ([18, Theorem 1.4])**.** *Assume that $\mu \in \mathcal{P}_2(\Omega)$ is absolutely continuous with respect to the Lebesgue measure. Then, there exists a unique optimal transport plan $\gamma^\star$ that realizes the infimum in (2) and it is of the form $(Id \times T)_{\#}\mu$, for a measurable function $T : \Omega \to \Omega$. Furthermore, there exists at least a Kantorovich potential ψ whose gradient $\nabla\psi$ is uniquely determined μ-almost everywhere. The function T and the potential ψ are linked by $T(x) = x - \nabla\psi(x)$.*

The measurable function $T : \Omega \to \Omega$ is referred to as the optimal transport map from μ to ν. This result implies that there exists a solution for transporting samples from μ to samples from ν and this solution is optimal in the sense that it minimizes the ℓ_2 displacement. However, identifying this solution is highly non-trivial. In the discrete case, effective solutions have been proposed [19]. However, for continuous and high-dimensional probability measures, constructing an actual transport plan remains a challenge. Even if recent contributions [20] have made it possible to rapidly compute $\mathcal{W}_2$, they do so without constructing the optimal map T, which is our objective here.

2.2 Wasserstein spaces and gradient flows

By [21, Proposition 7.1.5], $\mathcal{W}_2$ is a distance over $\mathcal{P}(\Omega)$. In addition, if $\Omega \subset \mathbb{R}^d$ is compact, the topology associated with $\mathcal{W}_2$ is equivalent to the weak convergence of probability measures and $(\mathcal{P}(\Omega), \mathcal{W}_2)$[2] is compact. The metric space $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$ is called the *Wasserstein space*.

In this study, we are interested in functional optimization problems in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$, such as $\min_{\mu \in \mathcal{P}_2(\Omega)} \mathcal{F}(\mu)$, where $\mathcal{F}$ is the functional that we would like to minimize. Similar to Euclidean spaces, one way to formulate this optimization problem is to construct a gradient flow of the form $\partial_t \mu_t = -\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)$, where $\nabla_{\mathcal{W}_2}$ denotes a notion of gradient in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$. If such a flow can be constructed, then one can utilize it both for practical algorithms and theoretical analysis.

Gradient flows $\partial_t \mu_t = \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)$ with respect to a functional $\mathcal{F}$ in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$ have strong connections with partial differential equations (PDE) that are of the form of a *continuity equation* [22]. Indeed, it is shown than under appropriate conditions on $\mathcal{F}$ (see e.g.[21]), $(\mu_t)_t$ is a solution of the gradient flow if and only if it admits a density ρ_t with respect to the Lebesgue measure for all $t \geq 0$, and solves the continuity equation given by: $\partial_t \rho_t + \operatorname{div}(v\rho_t) = 0$, where v denotes a vector field and div denotes the divergence operator. Then, for a given gradient flow in $(\mathcal{P}_2(\Omega), \mathcal{W}_2)$, we are interested in the evolution of the densities ρ_t, i.e. the PDEs which they solve. Such PDEs are of our particular interest since they have a key role for building practical algorithms.

2.3 Sliced-Wasserstein distance

In the one-dimensional case, i.e. $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, $\mathcal{W}_2$ has an analytical form, given as follows: $\mathcal{W}_2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(\tau) - F_\nu^{-1}(\tau)|^2 \, d\tau$, where F_μ and F_ν denote the cumulative distribution functions (CDF) of μ and ν, respectively, and F_μ^{-1}, F_ν^{-1} denote the inverse CDFs, also called quantile functions (QF). In this case, the optimal transport map from μ to ν has a closed-form formula as well, given as follows: $T(x) = (F_\nu^{-1} \circ F_\mu)(x)$ [7]. The optimal map T is also known as the *increasing arrangement*, lwhich maps each quantile of μ to the same quantile of ν, e.g. minimum to minimum, median to median, maximum to maximum. Due to Theorem 1, the derivative of the corresponding Kantorovich potential is given as $\psi'(x) \triangleq \partial_x \psi(x) = x - (F_\nu^{-1} \circ F_\mu)(x)$.

In the multidimensional case $d > 1$, building a transport map is much more difficult. The nice properties of the one-dimensional Wasserstein distance motivate the usage of *sliced-Wasserstein distance* ($\mathcal{SW}_2$) for practical applications. Before formally defining $\mathcal{SW}_2$, let us first define the orthogonal projection $\theta^*(x) \triangleq \langle \theta, x \rangle$ for any direction $\theta \in \mathbb{S}^{d-1}$ and $x \in \mathbb{R}^d$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner-product and $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ denotes the d-dimensional unit sphere. Then, the $\mathcal{SW}_2$ distance is formally defined as follows:

$$\mathcal{SW}_2(\mu, \nu) \triangleq \int_{\mathbb{S}^{d-1}} \mathcal{W}_2(\theta^*_\# \mu, \theta^*_\# \nu) \, d\theta, \tag{4}$$

where $d\theta$ represents the uniform probability measure on $\mathbb{S}^{d-1}$. As shown in [15], $\mathcal{SW}_2$ is indeed a distance metric and induces the same topology as $\mathcal{W}_2$ for compact domains.

[2]Note that in that case, $\mathcal{P}_2(\Omega) = \mathcal{P}(\Omega)$

The $\mathcal{SW}_2$ distance has important practical implications: provided that the distributions $\theta^*_\#\mu$ and $\theta^*_\#\nu$ can be computed, then for any $\theta \in \mathbb{S}^{d-1}$, the distance $\mathcal{W}_2(\theta^*_\#\mu, \theta^*_\#\nu)$, as well as its optimal transport map and the corresponding Kantorovich potential can be analytically computed (since the projected measures are one-dimensional). Therefore, one can easily approximate (4) by using a simple Monte Carlo scheme that draws uniform random samples from $\mathbb{S}^{d-1}$ and replace the integral in (4) with a finite-sample average. Thanks to its computational benefits, $\mathcal{SW}_2$ was very recently considered for OT-based VAEs and GANs [9, 23, 24], appearing as a stable alternative to the adversarial methods.

3 Regularized Sliced-Wasserstein Flows for Generative Modeling

3.1 Construction of the gradient flow

We propose in this paper to consider the minimization of the functional $\mathcal{F}^\nu_\lambda$ on $\mathcal{P}_2(\Omega)$, that is defined as follows:

$$\mathcal{F}^\nu_\lambda(\mu) \triangleq \frac{1}{2}\mathcal{SW}_2^2(\mu,\nu) + \lambda\mathcal{H}(\mu), \tag{5}$$

where $\lambda > 0$ is a regularization parameter and $\mathcal{H}$ denotes the negative entropy defined by $\mathcal{H}(\mu) \triangleq \int_\Omega \rho(x)\log\rho(x)dx$ if μ has density ρ with respect to the Lebesgue measure and $\mathcal{H}(\mu) = +\infty$ otherwise. Note that the case $\lambda = 0$ has been already proposed and studied in [15] in a more general OT context. Here, in order to introduce the necessary noise inherent to generative model, we suggest to penalize the slice-Wasserstein distance using $\mathcal{H}$. In other words, the main idea is to find a measure $\mu^\star$ that is close to ν as much as possible and also has a certain amount of entropy to make sure that it is sufficiently expressive for generative modeling purposes. The importance of the entropy regularization becomes prominent in practical applications where we have finitely many data samples that are assumed to be drawn from ν. In such a circumstance, the regularization would prevent $\mu^\star$ to collapse on the data points and therefore avoid 'over-fitting' to the data distribution. Note that this regularization is fundamentally different than the one used in Sinkhorn distances[25].

In the next result, we show that there exists a flow $(\mu_t)_{t\geq 0}$ in $(\mathcal{P}(\overline{\mathrm{B}}(0,r)), \mathcal{W}_2)$ which decreases along $\mathcal{F}^\nu_\lambda$, where $\overline{\mathrm{B}}(0,a)$ denotes the closed unit ball centered at 0 and radius a. This flow will be referred to as a generalized minimizing movement scheme (see Definition 1 in Appendix). In addition, the flow $(\mu_t)_{t\geq 0}$ admits a density ρ_t with respect to the Lebesgue measure for all $t > 0$ and $(\rho_t)_{t\geq 0}$ is solution of a non-linear PDE (in the weak sense).

Theorem 2. *Let ν be a probability measure on $\overline{\mathrm{B}}(0,1)$ with a strictly positive smooth density. Choose a regularization constant $\lambda > 0$ and radius $r > \sqrt{d}$. Assume that $\mu_0 \in \mathcal{P}(\overline{\mathrm{B}}(0,r))$ is absolutely continuous with respect to the Lebesgue measure with density $\rho_0 \in \mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$. There exists a generalized minimizing movement scheme $(\mu_t)_{t\geq 0}$ given by Theorem S2 in Appendixand if ρ_t stands for the density of μ_t for all $t \geq 0$, then $(\rho_t)_t$ satisfies the following continuity equation:*

$$\frac{\partial\rho_t}{\partial t} = -\operatorname{div}(v_t\rho_t) + \lambda\Delta\rho_t, \qquad v_t(x) \triangleq v(x,\mu_t) = -\int_{\mathbb{S}^{d-1}} \psi'_{t,\theta}(\langle x,\theta\rangle)\theta d\theta \tag{6}$$

in a weak sense. Here, Δ *denotes the Laplacian operator,* div *the divergence operator, and* $\psi_{t,\theta}$ *denotes the Kantorovich potential between* $\theta^*_{\#}\mu_t$ *and* $\theta^*_{\#}\nu$*.*

The precise statement of this Theorem, related results and its proof are postponed to Appendix. For its proof, we use the technique introduced in [26]: we first prove the existence of a generalized minimizing movement scheme by showing that the solution curve $(\mu_t)_t$ is a limit of the solution of a time-discretized problem. Then we prove that the curve $(\rho_t)_t$ solves the PDE given in (6).

3.2 Connection with stochastic differential equations

As a consequence of the entropy regularization, we obtain the Laplacian operator Δ in the PDE given in (6). We therefore observe that the overall PDE is a Fokker-Planck-type equation [27] that has a well-known probabilistic counterpart, which can be expressed as a stochastic differential equation (SDE). More precisely, let us consider a stochastic process $(X_t)_t$, that is the solution of the following SDE:

$$dX_t = v(X_t, \mu_t)dt + \sqrt{2\lambda}dW_t, \qquad X_0 \sim \mu_0 \tag{7}$$

where W_t denotes the standard Brownian motion. Then, the probability distribution of X_t at time t solves the PDE given in (6). This informally means that, if we could simulate (7), then the distribution of X_t would converge to the solution of (5), therefore, we could use the sample paths $(X_t)_t$ as samples drawn from $(\mu_t)_t$. However, in practice this is not possible due to two reasons: (i) the drift v_t cannot be computed analytically since it depends on the probability distribution of X_t, (ii) the SDE (7) is a continuous-time process, it needs to be discretized.

We now focus on the first issue. We observe that the SDE (7) is similar to McKean-Vlasov SDEs [28, 29], a family of SDEs whose drift depends on the distribution of X_t. By using this connection, we can borrow tools from the relevant SDE literature [30, 31] for developing an approximate simulation method for (7).

Our approach is based on defining a *particle system* that serves as an approximation to the original SDE (7). The particle system can be written as a collection of SDEs, given as follows [32]:

$$dX^i_t = v(X^i_t, \mu^N_t)dt + \sqrt{2\lambda}dW^i_t\,, \quad i = 1, \dots, N, \tag{8}$$

where i denotes the particle index, $N \in \mathbb{N}_+$ denotes the total number of particles, and $\mu^N_t = (1/N)\sum_{j=1}^N \delta_{X^j_t}$ denotes the empirical distribution of the particles $\{X^j_t\}_{j=1}^N$. This particle system is particularly interesting, since (i) one typically has $\lim_{N\to\infty} \mu^N_t = \mu_t$ with a rate of convergence of order $\mathcal{O}(1/\sqrt{N})$ for all t [30, 31], and (ii) each of the particle systems in (8) can be simulated by using an Euler-Maruyama discretization scheme. We note that the existing theoretical results in [28, 29] do not directly apply to our case due to the non-standard form of our drift. However, we conjecture that a similar result holds for our problem as well. Proving such a result would be very involved and it is out of the scope of this study.

3.3 Approximate Euler-Maruyama discretization

In order to be able to simulate the particle SDEs (8) in practice, we propose an approximate Euler-Maruyama discretization for each particle SDE. The algorithm iteratively applies the following update equation:

$$\bar{X}_0^i \overset{\text{i.i.d.}}{\sim} \mu_0, \qquad \bar{X}_{k+1}^i = \bar{X}_k^i + h\hat{v}_k(\bar{X}_k^i) + \sqrt{2\lambda h} Z_{k+1}^i, \qquad \forall i \in \{1, \dots, N\} \tag{9}$$

where $k \in \mathbb{N}_+$ denotes the iteration number, Z_k^i is a standard Gaussian random vector in $\mathbb{R}^d$, h denotes the step-size, and $\hat{v}_k$ is a short-hand notation for a computationally tractable estimator of the original drift $v(\cdot, \bar{\mu}_{kh}^N)$, with $\bar{\mu}_{kh}^N = (1/N)\sum_{j=1}^N \delta_{\bar{X}_k^j}$ being the empirical distribution of $\{\bar{X}_k^j\}_{j=1}^N$. A question of fundamental practical importance is how to compute this function $\hat{v}$.

We propose to approximate the integral in (6) via a simple Monte Carlo estimate. At each iteration k, this is done by drawing N_θ uniform i.i.d. samples from the sphere $\mathbb{S}^{d-1}$, $\{\theta_{k,n}\}_{n=1}^{N_\theta}$, and computing:

$$\hat{v}_k(x) \triangleq -(1/N_\theta)\sum_{n=1}^{N_\theta} \psi'_{k,\theta_{k,n}}(\langle \theta_{k,n}, x\rangle)\theta_{k,n}, \tag{10}$$

where for any θ, $\psi'_{k,\theta}$ is the derivative of the Kantorovich potential (cf. Section 2) that is applied to the OT problem from $\theta^*_\# \bar{\mu}_{kh}^N$ to $\theta^*_\# \nu$: i.e. $\psi'_{k,\theta}(z) = \left[z - (F^{-1}_{\theta^*_\#\nu} \circ F_{\theta^*_\# \bar{\mu}_{kh}^N})(z)\right]$.

For any particular $\theta \in \mathbb{S}^{d-1}$, the QF $F^{-1}_{\theta^*_\#\nu}$ for the projection of the target distribution ν on θ can be easily computed from the data. This is done by first computing the projections $\langle \theta, y_i\rangle$ for all data points y_i, and then computing the empirical quantile function for this set of P scalars. Similarly, $F_{\theta^*_\# \bar{\mu}_{kh}^N}$, the CDF of the particles at iteration k, is easy to compute: we first project all particles $\bar{X}_k^i$ to get $\langle \theta, \bar{X}_k^i\rangle$, and then compute the empirical CDF of this set of N scalar values.

In both cases, the true CDF and quantile functions are approximated as a linear interpolation between a set of the computed $Q \in \mathbb{N}_+$ empirical quantiles. Another source of approximation here comes from the fact that the target ν will in practice be a collection of Dirac measures on the observations y_i. Since it is currently common to have a very large datasets, we believe this approximation to be accurate in practice for the target.

Algorithm 1: Sliced-Wasserstein Flow (SWF)

input : $\mathcal{D} \equiv \{y_i\}_{i=1}^P$, μ_0, N, N_θ, h, λ
output : $\{\bar{X}_K^i\}_{i=1}^N$
// Initialize the particles
$\bar{X}_0^i \overset{\text{i.i.d.}}{\sim} \mu_0$, $\quad i = 1, \dots, N$
for $k = 0, \dots K-1$ **do**
  // Generate random directions
  $\theta_{k,n} \sim \text{Uniform}(\mathbb{S}^{d-1})$, $\quad n = 1, \dots, N_\theta$
  for $\theta \in \{\theta_{k,n}\}_{n=1}^{N_\theta}$ **do**
    // CDF of projected particles
    $F_{\theta^*_\# \bar{\mu}_{kh}^N} = \text{CDF}\{\langle \theta, \bar{X}_k^i\rangle\}_{i=1}^N$
    // Quantiles of projected target
    $F^{-1}_{\theta^*_\#\nu} = \text{QF}\{\langle \theta, y_i\rangle\}_{i=1}^P$
  // Update the particles
  $\bar{X}_{k+1}^i = \bar{X}_k^i - h\hat{v}_k(\bar{X}_k^i) + \sqrt{2\lambda h} Z_{k+1}^i$
  $i = 1, \dots, N$

Even though the error induced by these approximation schemes can be incorporated into our current analysis framework, we choose to neglect it for now, because (i) all these one-dimensional computations can be done very accurately and (ii) quantization of the empirical CDF and QF can be modeled as additive

Gaussian noise that enters our discretization scheme (9) [33]. Therefore, we will assume that $\hat{v}_k$ is an unbiased estimator of v, i.e. $\mathbb{E}[\hat{v}(x,\mu)] = v(x,\mu)$, for any x and μ, where the expectation is taken over $\theta_{k,n}$.

The overall algorithm is illustrated in Algorithm 1. It is remarkable that the updates of the particles only involves the learning data $\{y_i\}$ through the CDF of its projections on the many $\theta_{k,n} \in \mathbb{S}^{d-1}$. This has a fundamental consequence of high practical interest: these CDF may be computed in a massively distributed manner that is independent of the sliced Wasserstein flow. This aspect is reminiscent of the *compressive learning* methodology [34], except we exploit quantiles of random projections here, instead of random moments as done there.

Besides, we can obtain further reductions in the computing time if the CDF $F_{\theta^*_\# \nu}$ for the target is computed on random mini-batches of the data, instead of the whole dataset of size P. This simplified procedure also has some interesting consequences in privacy-preserving settings: since we can vary the number of projection directions N_θ for a each data point y_i, by using the compressed sensing theory [35], we may guarantee that y_i cannot be recovered via these projections, by simply picking fewer projections than necessary for reconstruction.

3.4 Finite-time analysis for the infinite particle regime

In this section we will analyze the behavior of the proposed algorithm in the asymptotic regime where the number of particles $N \to \infty$. Within this regime, we will assume that the original SDE (7) can be directly simulated by using an approximate Euler-Maruyama scheme, defined as follows:

$$\bar{X}_0 \overset{\text{i.i.d.}}{\sim} \mu_0, \qquad \bar{X}_{k+1} = \bar{X}_k + h\hat{v}(\bar{X}^i_k, \bar{\mu}_{kh}) + \sqrt{2\lambda h} Z_{k+1}, \tag{11}$$

where $\bar{\mu}_{kh}$ denotes the law of $\bar{X}_k$ with step size h and $\{Z_k\}_k$ denotes a collection of standard Gaussian random variables. Apart from its theoretical significance, this scheme is also practically relevant, since one would expect that it captures the behavior of the particle method (9) with large number of particles.

In practice, we would like to approximate the measure sequence $(\mu_t)_t$ as accurate as possible, where μ_t denotes the law of X_t. Therefore, we are interested in analyzing the distance $\|\bar{\mu}_{Kh} - \mu_T\|_{\text{TV}}$, where $T = Kh$ and $\|\mu - \nu\|_{\text{TV}}$ denotes the total variation distance between two probability measures μ and ν: $\|\mu - \nu\|_{\text{TV}} \triangleq \sup_{A \in \mathcal{B}(\Omega)} |\mu(A) - \nu(A)|$.

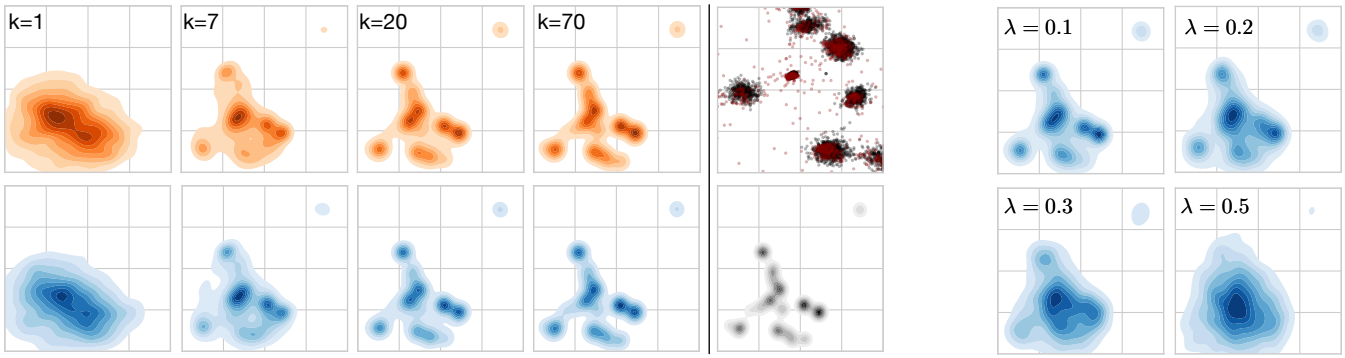
In order to analyze this distance, we exploit the connections between (11) and the stochastic gradient Langevin dynamics (SGLD) algorithm [16], which is a Bayesian posterior sampling method, and is obtained as a discretization of an SDE whose drift has a much simpler form. We then bound the distance by extending the recent results on SGLD [17] to time- and measure-dependent drifts, that are of our interest in the paper.

We now present our second main theoretical result. We present all our assumptions and the explicit forms of the constants in Appendix.

Theorem 3. *Assume that the conditions given in Appendixhold. Then, the following bound holds for $T = Kh$:*

$$\|\bar{\mu}_{Kh} - \mu_T\|^2_{\text{TV}} \leq \delta_\lambda \left\{ \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} \right\}, \tag{12}$$

for some $C_1, C_2, L > 0$, $\delta \in (0,1)$, and $\delta_\lambda > 1$.



(a) (b)

Figure 1: a) **Left:** Distribution of particles (contour plots) during the estimation (top) and prediction (bottom) stages. **Right:** (top) Close-up of some generated particles in red superimposed with data points in black. (bottom) Target distribution. b) Influence of the regularization parameter λ.

Here, the constants C_1, C_2, L are related to the regularity and smoothness of the functions v and $\hat{v}$, δ is directly proportional to the variance of $\hat{v}$, and δ_λ is inversely proportional to λ. The theorem shows that if we choose h small enough, we can have a non-asymptotic error guarantee, which is formally shown in the following corollary.

Corollary 1. *Assume that the conditions of Theorem 3 hold. Then for all $\varepsilon > 0$, $K \in \mathbb{N}_+$, setting $h = (3/C_1) \wedge \left(\frac{2\varepsilon^2\lambda}{\delta_\lambda L^2 T}(1+3\lambda d)^{-1}\right)^{1/2}$, we have*

$$\|\bar{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}} \leq \varepsilon + \left(\frac{C_2\delta_\lambda\delta Kh}{4\lambda}\right)^{1/2} \tag{13}$$

for $T = Kh$.

This corollary shows that for a large horizon T, the approximate drift $\hat{v}$ should have a small variance in order to obtain accurate estimations. This result is similar to Theorem 2.1 of [17]: for small ε the variance of the approximate drift should be small as well. On the other hand, we observe that the error decreases as λ increases. This behavior is expected since for large λ, the Brownian term in (7) dominates the drift, which makes the simulation easier.

4 Experiments

In this section, we evaluate the SWF algorithm on both synthetic and real data settings. In all cases, the initial distribution μ_0 is selected as the standard Gaussian distribution on $\mathbb{R}^d$, we take $Q = 100$ quantiles, which proved sufficient to approximate the quantile functions, and we have observed that $N = 3000$ particles are sufficient. We provide an example implementation in Appendix.

4.1 Gaussian Mixture Model

We perform the first set of experiments on synthetic data where we consider a standard Gaussian mixture model (GMM). We set the number of the mixture components to 20 and for each component we randomly draw the weight, covariance matrix and the centroid. We make sure that the centroids are sufficiently distant from each other in order to make the problem more challenging. Given the model parameters, we generate $P = 50000$ data samples in each experiment.

In our first experiment, we set $d = 2$ for visualization purposes and illustrate the general behavior of the algorithm. Figure 1(a) shows the evolution of the particles through the iterations. Here, we set $N_\theta = 30$, $h = 1$, and $\lambda = 10^{-4}$. We observe that the empirical distribution of the particles converges rapidly to the target distribution. Furthermore, we can see that the QF, $F^{-1}_{\theta^*_\# \bar{\mu}^N_{kh}}$ that is computed with one set of particles (so-called the *estimation* stage) can be perfectly re-used for new unseen particles in a subsequent *prediction* stage. In both cases, we observe two remarkable outcomes: (i) Even when some modes are isolated from the others, SWF is able to capture them successfully and we never observe a mode collapse. This is due to the OT nature of the procedure. (ii) The generated particles do not collapse on the data points, thanks to the entropy regularization.

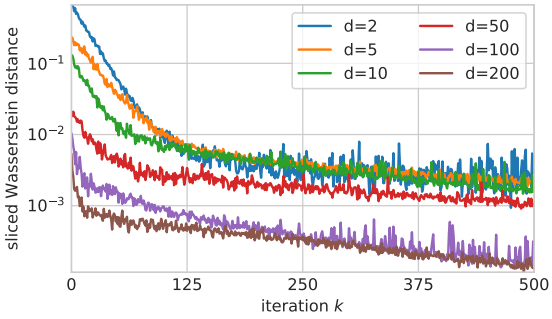


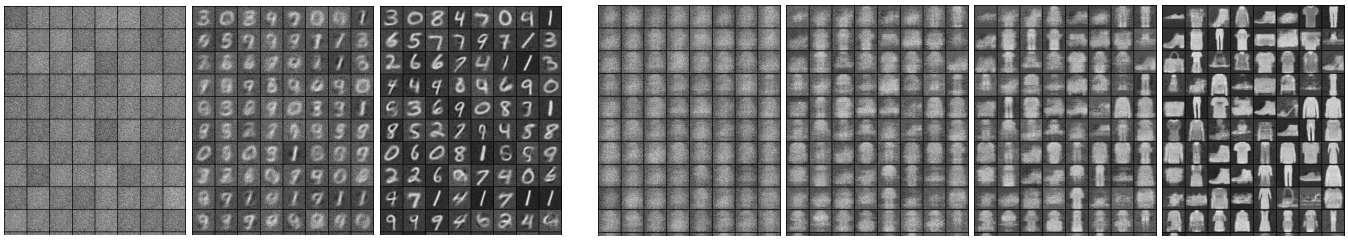
Figure 2: Approximately computed $\mathcal{SW}_2$ between the output $\bar{\mu}^N_k$ and data distribution ν in the GMM model for different data dimensions d.

In our second experiment, we investigate the effect of the level of the regularization. We use the same setting as the previous experiment, whereas we differ the value of λ and run the algorithm for sufficiently many iterations. As we can observe from Figure 1(b), the distribution of the particles becomes more spread with increasing λ. This is due to the increment of the entropy, as expected.

We also illustrate the behavior of the algorithm for varying dimensionality d. Since visualizing the results becomes non-trivial for large d, in this experiment we directly monitor the (approximately computed) $\mathcal{SW}_2$ distance between the distribution of the particles and the data distribution. Even though minimizing this distance is not the real objective of our method, arguably, it is still a good proxy for understanding the convergence behavior. Figure 2 illustrates the results. We observe that, for all choices of d, we see a steady decrease in the cost for all runs, which is in line with our theory. We also observe that the magnitude of $\mathcal{SW}_2$ decreases as d increases. This outcome can be explained by the fact that $\mathcal{SW}_2 = \mathcal{O}(d^{-1/2}\mathcal{W}_2)$ (cf. [15]).

4.2 Experiments on real data

In a second set of experiments, we test the SWF algorithm on two real datasets. (i) The traditional MNIST dataset that contains 70K binary images corresponding to different digits (of size 28×28, i.e. $d = 784$). (ii) The recently proposed FashionMNIST dataset [36], that contains 50000 gray-scale images. All images were interpolated as 64×64, yielding $d = 4096$. This dataset is advocated as more challenging than MNIST.



(a) MNIST (b) Fashion MNIST

Figure 3: The evolution of the particles through 15000 iterations on different datasets.

Our goal in these experiments is to capture the structure of the data distribution such that the particles that are generated by the algorithm will be samples from this unknown data distribution. In these experiments, we set $\lambda = 10^{-6}$ and $N_\theta = 200$. We will present visual results for qualitative inspection. More results with higher resolution are given in Appendix.

Figures 3(a) and 3(b), show that SWF is able to generate samples from the datasets in a few thousand iterations. We can observe that, the generated samples for the MNIST dataset are considerably accurate. For the FashionMNIST dataset, the samples capture the prominent features of the training samples; the generated samples take the form of various clothings along the iterations. By considering that SWF only requires the projections of the data points, in a way, all these samples are generated without seeing the actual dataset.

Another important advantage of SWF is its low computational requirements. The whole experiment on the FashionMNIST requires around 1 hour of computational time on the CPU of a standard laptop computer, to be compared with the significant resource requirements of the current IGM methods.

5 Conclusion and Future Directions

In this study, we proposed SWF, a theoretically grounded nonparametric algorithm for efficient implicit generative modeling. Our approach lies in the intersection of OT, Wasserstein gradient flows, and SDEs. This connection allowed us to convert the IGM problem to a non-linear SDE simulation problem. We provided finite-time error bounds for the infinite-particle regime and established explicit links between the algorithm parameters and the overall error. We conducted experiments on both synthetic and real datasets, where we showed that the results support our theory: SWF is able to generate samples from challenging distributions with low computational requirements.

The SWF algorithm opens up interesting future directions: (i) extension of the algorithm to differentially private settings [37] by exploiting the fact that it only requires random projections of the data, (ii) showing the convergence scheme of the particle system (8) to the original SDE (7), (iii) providing error bounds directly for the particle scheme (9), (iv) combining SWF with existing IGM approaches in order to be able to simulate the particles in a lower dimensional space.

Acknowledgments

This work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014) and KAMoulox (ANR-15-CE38-0003-01) projects, and by the industrial chair Machine Learning for Big Data from Télécom ParisTech.

References

[1] Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

[2] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[6] S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5551–5559, 2017.

[7] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[8] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[9] S. Kolouri, C. E. Martin, and G. K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[12] X. Guo, J. Hong, T. Lin, and N. Yang. Relaxed Wasserstein with applications to GANs. *arXiv preprint arXiv:1705.07164*, 2017.

[13] N. Lei, K. Su, L. Cui, S.-T. Yau, and D. X. Gu. A geometric view of optimal transportation and generative model. *arXiv preprint arXiv:1710.05488*, 2017.

[14] A. Genevay, G. Peyré, and M. Cuturi. Gan and vae from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

[15] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

[16] M. Welling and Y. W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

[17] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703, 2017.

[18] F. Santambrogio. Introduction to optimal transport theory. *arXiv preprint arXiv:1009.3856*, 2010.

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[20] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

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

[22] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[23] Ishan Deshpande, Ziyu Zhang, and Alexander Schwing. Generative modeling using the sliced wasserstein distance. *arXiv preprint arXiv:1803.11188*, 2018.

[24] J. Wu, Z. Huang, W. Li, and L. V. Gool. Sliced wasserstein generative models. *arXiv preprint arXiv:1706.02631*, abs/1706.02631, 2018.

[25] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[26] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[27] V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov. *Fokker-Planck-Kolmogorov Equations*, volume 207. American Mathematical Soc., 2015.

[28] A Y. Veretennikov. On ergodic measures for mckean-vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 471–486. Springer, 2006.

[29] Y. S. Mishura and A. Y. Veretennikov. Existence and uniqueness theorems for solutions of mckean–vlasov stochastic equations. *arXiv preprint arXiv:1603.02212*, 2016.

[30] F. Malrieu. Convergence to equilibrium for granular media equations and their Euler schemes. *Ann. Appl. Probab.*, 13(2):540–560, 2003.

[31] P. Cattiaux, A. Guillin, and F. Malrieu. Probabilistic approach for granular media equations in the non uniformly convex case. *Prob. Theor. Rel. Fields*, 140(1-2):19–40, 2008.

[32] M. Bossy and D. Talay. A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Mathematics of Computation of the American Mathematical Society*, 66(217):157–192, 1997.

[33] A. W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

[34] Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, 2017.

[35] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[36] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[37] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[38] A. 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.

Appendix

6 Proof of Theorem 2

We first need to generalize [15, Lemma 5.4.3] to distribution $\rho \in \mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, $r > 0$.

Theorem 4. *Let ν be a probability measure on $\overline{\mathrm{B}}(0,1)$ with a strictly positive smooth density. Fix a time step $h>0$, regularization constant $\lambda>0$ and a radius $r>\sqrt{d}$. For any probability measure μ_0 on $\overline{\mathrm{B}}(0,r)$ with density $\rho_0 \in \mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, there is a probability measure μ on $\overline{\mathrm{B}}(0,r)$ minimizing:*

$$\mathcal{G}(\mu) = \mathcal{F}^\nu_\lambda(\mu) + \frac{1}{2h}\mathcal{W}_2^2(\mu,\mu_0),$$

where $\mathcal{F}^\nu_\lambda$ is given by (5). Moreover the optimal μ has a density ρ on $\overline{\mathrm{B}}(0,r)$ and:

$$||\rho||_{\mathrm{L}^\infty} \leq (1+h/\sqrt{d})^d ||\rho_0||_{\mathrm{L}^\infty}. \tag{14}$$

Proof. The set of measures supported on $\overline{\mathrm{B}}(0,r)$ is compact in the topology given by $\mathcal{W}_2$ metric. Furthermore by [21, Lemma 9.4.3] $\mathcal{H}$ is lower semicontinuous on $(\mathcal{P}(\overline{\mathrm{B}}(0,r)),\mathcal{W}_2)$. Since by [15, Proposition 5.1.2, Proposition 5.1.3], $\mathcal{SW}_2$ is a distance on $\mathcal{P}(\overline{\mathrm{B}}(0,r))$, dominated by $d^{-1/2}\mathcal{W}_2$, we have:

$$|\mathcal{SW}_2(\pi_0,\nu) - \mathcal{SW}_2(\pi_1,\nu)| \leq \mathcal{SW}_2(\pi_0,\pi_1) \leq \frac{1}{\sqrt{d}}\mathcal{W}_2(\pi_0,\pi_1).$$

The above means that $\mathcal{SW}_2(\cdot,\nu)$ is continuous with respect to topology given by $\mathcal{W}_2$, which implies that $\mathcal{SW}_2^2(\cdot,\nu)$ is continuous in this topology as well. Therefore $\mathcal{G}:\mathcal{P}(\overline{\mathrm{B}}(0,r)) \to (-\infty,+\infty]$ is a lower semicontinuous function on the compact set $(\mathcal{P}(\overline{\mathrm{B}}(0,r)),\mathcal{W}_2)$. Hence there exists a minimum μ of $\mathcal{G}$ on $\mathcal{P}(\overline{\mathrm{B}}(0,r))$. Furthermore, since $\mathcal{H}(\pi)=+\infty$ for measures π that do not admit a density with respect to Lebesgue measure, the measure μ must admit a density ρ.

If ρ_0 is smooth and positive on $\overline{\mathrm{B}}(0,r)$, the inequality 14 is true by [15, Lemma 5.4.3.] When ρ_0 is just in $\mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, we proceed by smoothing. For $t\in(0,1]$, let ρ_t be a function obtained by convolution of ρ_0 with a Gaussian kernel $(t,x,y)\mapsto (2\pi)^{d/2}\exp(\|x-y\|^2/2)$, restricting the result to $\overline{\mathrm{B}}(0,r)$ and normalizing to obtain a probability density. Then $(\rho_t)_t$ are smooth positive densities, and it is easy to see that $\lim_{t\to 0}||\rho_t||_{\mathrm{L}^\infty} \leq ||\rho_0||_{\mathrm{L}^\infty}$. Furthermore, if we denote by μ_t the measure on $\overline{\mathrm{B}}(0,r)$ with density ρ_t, then μ_t converge weakly to μ_0. For $t\in(0,1]$ let $\hat{\mu}_t$ be the minimum of $\mathcal{F}^\nu_\lambda(\cdot)+\frac{1}{2h}\mathcal{W}_2^2(\cdot,\mu_t)$, and let $\hat{\rho}_t$ be the density of $\hat{\mu}_t$. Using [15, Lemma 5.4.3.] we get

$$||\hat{\rho}_t||_{\mathrm{L}^\infty} \leq (1+h\sqrt{d})^d ||\rho_t||_{\mathrm{L}^\infty}.$$

so $\hat{\rho}_t$ lies in a ball of finite radius in L^∞. Using compactness of $\mathcal{P}(\overline{\mathrm{B}}(0,r))$ in weak topology and compactness of closed ball in $\mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$ in weak star topology, we can choose a subsequence $\hat{\mu}_{t_k},\hat{\rho}_{t_k}$, $\lim_{k\to+\infty}t_k=0$, that converges along that subsequence to limits $\hat{\mu}$, $\hat{\rho}$. Obviously $\hat{\rho}$ is the density of $\hat{\mu}$, since for any continuous function f on $\overline{\mathrm{B}}(0,r)$ we have:

$$\int \hat{\rho} f dx = \lim_{k\to\infty}\int \rho_{t_k} f dx = \lim_{k\to\infty}\int f d\mu_{t_k} = \int f d\mu.$$

Furthermore, since $\hat{\rho}$ is the weak star limit of a bounded subsequence, we have:

$$||\hat{\rho}||_{\mathrm{L}^\infty} \leq \limsup_{k\to\infty}(1+h\sqrt{d})^d ||\rho_{t_k}||_{\mathrm{L}^\infty} \leq (1+h\sqrt{d})^d ||\rho_0||_{\mathrm{L}^\infty}.$$

To finish, we just need to prove that $\hat{\mu}$ is a minimum of $\mathcal{G}$. We remind our reader, that we already established existence of some minimum μ (that might be different from $\hat{\mu}$). Since $\hat{\mu}_{t_k}$ converges weakly to $\hat{\mu}$ in $\mathcal{P}(\overline{\mathrm{B}}(0,r))$, it implies convergence in $\mathcal{W}_2$ as well since $\overline{\mathrm{B}}(0,r)$ is compact. Similarly μ_{t_k} converges to μ_0 in $\mathcal{W}_2$. Using the lower semicontinuity of $\mathcal{G}$ we now have:

$$\begin{aligned}\mathcal{F}^\nu_\lambda(\hat{\mu}) + \frac{1}{2h}\mathcal{W}_2^2(\hat{\mu},\mu_0) &\leq \liminf_{k\to\infty}\left(\mathcal{F}^\nu_\lambda(\hat{\mu}_{t_k}) + \frac{1}{2h}\mathcal{W}_2^2(\hat{\mu}_{t_k},\mu_0)\right)\\ &\leq \liminf_{k\to\infty}\mathcal{F}^\nu_\lambda(\mu) + \frac{1}{2h}\mathcal{W}_2^2(\mu,\mu_{t_k})\\ &+ \frac{1}{2h}\mathcal{W}_2^2(\hat{\mu}_{t_k},\mu_0) - \frac{1}{2h}\mathcal{W}_2^2(\hat{\mu}_{t_k},\mu_{t_k})\\ &= \mathcal{F}^\nu_\lambda(\mu) + \frac{1}{2h}\mathcal{W}_2^2(\mu,\mu_0),\end{aligned}$$

where the second inequality comes from the fact, that $\hat{\mu}_{t_k}$ minimizes $\mathcal{F}^\nu_\lambda(\cdot) + \frac{1}{2h}\mathcal{W}_2^2(\cdot,\mu_{t_k})$. From the above inequality and previously established facts, it follows that $\hat{\mu}$ is a minimum of $\mathcal{G}$ with density satisfying 14. □

Definition 1. ***Minimizing movement scheme*** *Let $r>0$ and $\mathcal{F}:\mathbb{R}_+\times\mathcal{P}(\overline{\mathrm{B}}(0,r))\times\mathcal{P}(\overline{\mathrm{B}}(0,r))\to\mathbb{R}$ be a functional. Let $\mu_0\in\mathcal{P}(\overline{\mathrm{B}}(0,r))$ be a starting point. For $h>0$ a piecewise constant trajectory $\mu^h:[0,\infty)\to\mathcal{P}(\overline{\mathrm{B}}(0,r))$ for $\mathcal{F}$ starting at μ_0 is a function such that:*

- $\mu^h(0)=\mu_0$.
- *μ^h is constant on each interval $[nh,(n+1)h)$, so $\mu^h(t)=\mu^h(nh)$ with $n=\lfloor t/h\rfloor$.*
- *$\mu^h((n+1)h)$ minimizes the functional $\zeta\mapsto\mathcal{F}(h,\zeta,\mu^h(nh))$, for all $n\in\mathbb{N}$.*

We say $\hat{\mu}$ is a minimizing movement scheme for $\mathcal{F}$ starting at μ_0, if there exists a family of piecewise constant trajectory $(\mu^h)_{h>0}$ for $\mathcal{F}$ such that $\hat{\mu}$ is a pointwise limit of μ^h as h goes to 0, i.e. for all $t\in\mathbb{R}_+$, $\lim_{h\to0}\mu^h(t)=\mu(t)$ in $\mathcal{P}(\overline{\mathrm{B}}(0,r))$. We say that $\tilde{\mu}$ is a generalized minimizing movement for $\mathcal{F}$ starting at μ_0, if there exists a family of piecewise constant trajectory $(\mu^h)_{h>0}$ for $\mathcal{F}$ and a sequence $(h_n)_n$, $\lim_{n\to\infty}h_n=0$, such that μ^{h_n} converges pointwise to $\tilde{\mu}$.

Theorem 5. *Let ν be a probability measure on $\overline{\mathrm{B}}(0,1)$ with a strictly positive smooth density. Fix a regularization constant $\lambda>0$ and radius $r>\sqrt{d}$. Given an absolutely continuous measure $\mu_0\in\mathcal{P}(\overline{\mathrm{B}}(0,r))$ with density $\rho_0\in\mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, there is a generalized minimizing movement scheme $(\mu_t)_t$ in $\mathcal{P}(\overline{\mathrm{B}}(0,r))$ starting from μ_0 for the functional defined by*

$$\mathcal{F}^\nu(h,\mu_+,\mu_-) = \mathcal{F}^\nu_\lambda(\mu_+) + \frac{1}{2h}\mathcal{W}_2^2(\mu_+,\mu_-). \tag{15}$$

Moreover for any time $t>0$, the probability measure $\mu_t=\mu(t)$ has density ρ_t with respect to the Lebesgue measure and:

$$||\rho_t||_{\mathrm{L}^\infty} \leq e^{dt\sqrt{d}}||\rho_0||_{\mathrm{L}^\infty}. \tag{16}$$

Proof. We start by noting, that by 4 for any $h > 0$ there exists a piecewise constant trajectory μ^h for 15 starting at μ_0. Furthermore for $t \geq 0$ measure $\mu_t^h = \mu^h(t)$ has density ρ_t^h, and:

$$||\rho_t^h||_{\mathrm{L}^\infty} \leq e^{d\sqrt{d}(t+h)}||\rho_0||_{\mathrm{L}^\infty}. \tag{17}$$

Let us choose $T > 0$. We denote $\rho^h(t,x) = \rho_t^h(x)$. For $h \leq 1$, the functions ρ^h lie in a ball in $\mathrm{L}^\infty([0,T]\times\overline{\mathrm{B}}(0,r))$, so from Banach-Alaoglu theorem there is a sequence h_n converging to 0, such that ρ^{h_n} converges in weak-star topology in $\mathrm{L}^\infty([0,T]\times\overline{\mathrm{B}}(0,r))$ to a certain limit ρ. Since ρ has to be nonnegative except for a set of measure zero, we assume ρ is nonnegative. We denote $\rho_t(x) = \rho(t,x)$. We will prove that for almost all t, ρ_t is a probability density and $\mu_t^{h_n}$ converges in $\mathcal{W}_2$ to a measure μ_t with density ρ_t.

First of all, for almost all $t \in [0,T]$, ρ_t is a probability density, since for any Borel set $A \subseteq [0,T]$ the indicator of set $A \times \overline{\mathrm{B}}(0,r)$ is integrable, and hence by definition of the weak-star topology:

$$\int_A \int_{\overline{\mathrm{B}}(0,r)} \rho_t(x)dxdt = \lim_{n\to\infty} \int_A \int_{\overline{\mathrm{B}}(0,r)} \rho_t^{h_n}(x)dxdt,$$

and so we have to have $\int \rho_t(x)dx = 1$ for almost all $t \in [0,T]$. Nonnegativity of ρ_t follows from nonnegativity of ρ.

We will now prove, that for almost all $t \in [0,T]$ the measures $\mu_t^{h_n}$ converge to a measure with density ρ_t. Let $t \in (0,T)$, take $\delta < \min(T-t, t)$ and $\zeta \in \mathrm{C}^1(\overline{\mathrm{B}}(0,r))$. We have:

$$\begin{aligned}\left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n} - \int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_m}\right| \leq \\ \left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n} - \frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_n} ds\right| + \left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_m} - \frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_m} ds\right| + \\ \left|\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_m} ds - \frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_n} ds\right|. \end{aligned} \tag{18}$$

Because $\mu_t^{h_n}$ have densities $\rho_t^{h_n}$ and both ρ^{h_n}, ρ^{h_m} converge to ρ in weak-star topology, the last element of the sum on the right hand side converges to zero, as $n, m \to \infty$. Next, we get a bound on the other two terms.

First, if we denote by γ the optimal transport plan between $\mu_t^{h_n}$ and $\mu_s^{h_n}$, we have:

$$\begin{aligned}\left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n} - \int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_n}\right|^2 \leq \int_{\overline{\mathrm{B}}(0,r)\times\overline{\mathrm{B}}(0,r)} |\zeta(x)-\zeta(y)|^2 \, d\gamma(x,y) \\ \leq ||\nabla\zeta||_\infty^2 \mathcal{W}_2^2(\mu_t^{h_n}, \mu_s^{h_n}). \end{aligned} \tag{19}$$

In addition, for $n_t = \lfloor t/h_n \rfloor$ and $n_s = \lfloor s/h_n \rfloor$ we have $\mu_t^{h_n} = \mu_{n_t h_n}^{h_n}$ and $\mu_s^{h_n} = \mu_{n_s h_n}^{h_n}$. For all $k \geq 0$ we have:

$$\mathcal{W}_2^2(\mu_{kh_n}^{h_n}, \mu_{(k+1)h_n}^{h_n}) \leq 2h_n(\mathcal{F}_\lambda^\nu(\mu_{kh_n}^{h_n}) - \mathcal{F}_\lambda^\nu(\mu_{(k+1)h_n}^{h_n}). \tag{20}$$

Using this result and (19) and assuming without loss of generality $n_t \leq n_s$, from the Cauchy-Schwartz inequality we get:

$$\begin{aligned}\mathcal{W}_2^2(\mu_t^{h_n}, \mu_s^{h_n}) &\leq \left(\sum_{k=n_t}^{n_s-1} \mathcal{W}_2(\mu_{kh_n}^{h_n}, \mu_{(k+1)h_n}^{h_n})\right)^2 \\ &\leq |n_t - n_s| \sum_{k=n_t}^{n_s 1} \mathcal{W}_2^2(\mu_{kh_n}^{h_n}, \mu_{(k+1)h_n}^{h_n}) \\ &\leq 2h_n|n_t - n_s|(\mathcal{F}_\lambda^\nu(\mu_{n_t h_n}^{h_n}) - \mathcal{F}_\lambda^\nu(\mu_{n_s h_n}^{h_n})) \leq 2C(|t-s| + h_n), \end{aligned} \tag{21}$$

where we used for the last inequality, denoting $C = \mathcal{F}_\lambda^\nu(\mu_0) - \min_{\mathcal{P}(\overline{\mathrm{B}}(0,r))} \mathcal{F}_\lambda^\nu$, that $(\mathcal{F}_\lambda^\nu(\mu_{kh_n}^{h_n}))_n$ is non-increasing by (20) and $\min_{\mathcal{P}(\overline{\mathrm{B}}(0,r))} \mathcal{F}_\lambda^\nu$ is finite since $\mathcal{F}_\lambda^\nu$ is lower semi-continuous. Finally, using Jensen's inequality, the above bound and 19 we get:

$$\begin{aligned}\left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n} - \frac{1}{2\delta}\int_{t-\delta}^{t+\delta} \int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_n} ds\right|^2 &\leq \frac{1}{2\delta}\int_{t-\delta}^{t+\delta} \left|\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n} - \int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_s^{h_n}\right|^2 ds \\ &\leq \frac{C||\nabla\zeta||_\infty^2}{\delta}\int_{t-\delta}^{t+\delta} (|t-s| + h_n) ds \\ &\leq 2C||\nabla\zeta||_\infty^2 (h_n + \delta).\end{aligned}$$

Together with (18), when taking $\delta = h_n$, this result means that $\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n}$ is a Cauchy sequence for all $t \in (0,T)$. On the other hand, since ρ^{h_n} converges to ρ in weak-star topology on L^∞, the limit of $\int_{\overline{\mathrm{B}}(0,r)} \zeta d\mu_t^{h_n}$ has to be $\int_{\overline{\mathrm{B}}(0,r)} \zeta(x)\rho_t(x)dx$ for almost all $t \in (0,T)$. This means that for almost all $t \in [0,T]$ sequence $\mu_t^{h_n}$ converges to a measure μ_t with density ρ_t.

Let $S \in [0,T]$ be the set of times such that for $t \in S$ sequence $\mu_t^{h_n}$ converges to μ_t. As we established almost all points from $[0,T]$ belong to S. Let $t \in [0,T] \setminus S$. Then, there exists a sequence of times $t_k \in S$ converging to t, such that μ_{t_k} converge to some limit μ_t. We have:

$$\mathcal{W}_2(\mu_t^{h_n}, \mu_t) \leq \mathcal{W}_2(\mu_t^{h_n}, \mu_{t_k}^{h_n}) + \mathcal{W}_2(\mu_{t_k}^{h_n}, \mu_{t_k}) + \mathcal{W}_2(\mu_{t_k}, \mu_t).$$

From which we have for all $k \geq 1$:

$$\limsup_{n\to\infty} \mathcal{W}_2(\mu_t^{h_n}, \mu_t) \leq \mathcal{W}_2(\mu_{t_k}, \mu_t) + \limsup_{n\to\infty} \mathcal{W}_2(\mu_t^{h_n}, \mu_{t_k}^{h_n}),$$

and using (21), we get $\mu_t^{h_n} \to \mu_t$. Furthermore, the measure μ_t has to have density, since $\rho_t^{h_n}$ lie in a ball in $\mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, so we can choose a subsequence of $\rho_t^{h_n}$ converging in weak-star topology to a certain limit $\hat{\rho}_t$, which is the density of μ_t.

We use now the diagonal argument to get convergence for all $t > 0$. Let $(T_k)_{k=1}^\infty$ be a sequence of times increasing to infinity. Let h_n^1 be a sequence converging to 0, such that $\mu_t^{h_n^1}$ converge to μ_t for all $t \in [0, T_1]$.

Using the same arguments as above, we can choose a subsequence h_n^2 of h_n^1, such that $\mu_t^{h_n^2}$ converges to a limit μ_t for all $t \in [0, T_2]$. Inductively, we construct subsequences h_n^k, and in the end take $h_n = h_n^n$. For this subsequence we have that $\mu_t^{h_n}$ converges to μ_t for all $t > 0$, and μ_t has a density satisfying the bound from the statement of the theorem.

Finally, note that (5) follows from (17). □

Theorem 6. *Let $(\mu_t)_{t\geq 0}$ be a generalized minimizing movement scheme given by Theorem 5 with initial distribution μ_0 with density $\rho_0 \in \mathrm{L}(\overline{\mathrm{B}}(0,r))$. We denote by ρ_t the density of μ_t for all $t \geq 0$. Then ρ_t satisfies the continuity equation:*

$$\frac{\partial \rho_t}{\partial t} + \operatorname{div}(v_t \rho_t) + \lambda \Delta \rho_t = 0\,, \qquad v_t(x) = -\int_{\mathbb{S}^{d-1}} \psi'_{t,\theta}(\langle x, \theta\rangle)\theta d\theta,$$

in a weak sense, that is for all $\xi \in \mathrm{C}_c^\infty([0,\infty) \times \overline{\mathrm{B}}(0,r))$ we have:

$$\int_0^\infty \int_{\overline{\mathrm{B}}(0,r)} \left[\frac{\partial \xi}{\partial t}(t,x) - v_t \nabla \xi(t,x) - \lambda \Delta \xi(t,x)\right] \rho_t(x) dx dt = -\int_{\overline{\mathrm{B}}(0,r)} \xi(0,x)\rho_0(x) dx.$$

Proof. Our proof is based on the proof of [15, Theorem 5.6.1]. We proceed in five steps.

(1) Let $h_n \to 0$ be a sequence given by Theorem 5, such that $\mu_t^{h_n}$ converges to μ_t pointwise. Furthermore we know that μ^{h_n} have densities ρ^{h_n} that converge to ρ in L^r, for $r \geq 1$, and in weak-star topology in L^∞. Let $\xi \in \mathrm{C}_c^\infty([0,\infty) \times \overline{\mathrm{B}}(0,r))$. We denote $\xi_k^n(x) = \xi(kh_n, x)$. Using part 1 of the proof of [15, Theorem 5.6.1], we obtain:

$$\int_{\overline{\mathrm{B}}(0,r)} \xi(0,x)\rho_0(x)dx + \int_0^\infty \int_{\overline{\mathrm{B}}(0,r)} \frac{\partial \xi}{\partial t}(t,x)\rho_t(x)dxdt \\ = \lim_{n\to\infty} -h_n \sum_{k=1}^\infty \int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n}(x) - \rho_{(k-1)h_n}^{h_n}(x)}{h_n} dx. \quad (22)$$

(2) Again, this part is the same as part 2 of the proof of [15, Theorem 5.6.1]. For any $\theta \in \mathbb{S}^{d-1}$ we denote by $\psi_{t,\theta}$ the unique Kantorovich potential from $\theta^*_\# \mu_t$ to $\theta^*_\# \nu$, and by $\psi_{t,\theta}^{h_n}$ the unique Kantorovich potential from $\theta^*_\# \mu_t^{h_n}$ to $\theta^*_\# \nu$. Then, by the same reasoning as part 2 of the proof of [15, Theorem 5.6.1], we get:

$$\int_0^\infty \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} (\psi_{t,\theta})'(\langle \theta, x\rangle)\langle \theta, \nabla \xi(x,t)\rangle d\theta d\mu_t(x) dt \\ = \lim_{n\to\infty} h_n \sum_{k=1}^\infty \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} \psi_{kh_n,\theta}^{h_n}(\theta^*)\langle \theta, \nabla \xi_k^n\rangle d\theta d\mu_{kh_n}^{h_n}. \quad (23)$$

(3) Since ξ is compactly supported and smooth, $\Delta\xi$ is Lipschitz, and so for any $t \geq 0$ if we take $k = \lfloor t/h_n \rfloor$ we get $|\Delta\xi_k^n(x) - \Delta\xi(t,x)| \leq Ch_n$ for some constant C. Let $T > 0$ be such that $\xi(t,x) = 0$ for $t > T$. We have:

$$\left| \sum_{k=1}^{\infty} h_n \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi_k^n(x) \rho_{kh_n}^{h_n}(x) dx - \int_0^{+\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi(t,x) \rho_t^{h_n}(x) dx dt \right| \leq CTh_n.$$

On the other hand, we know, that ρ^{h_n} converges to ρ in weak star topology on $\mathrm{L}^\infty([0,T] \times \overline{\mathrm{B}}(0,r))$, and $\Delta\xi$ is bounded, so:

$$\lim_{n \to +\infty} \left| \int_0^{+\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi(t,x) \rho_t^{h_n}(x) dx dt - \int_0^{+\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi(t,x) \rho_t(x) dx dt \right| = 0.$$

Combining those two results give:

$$\lim_{n \to \infty} h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi_k^n(x) \rho_{kh_n}^{h_n}(x) dx = \int_0^{+\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi(t,x) \rho_t(x) dx dt. \tag{24}$$

(4) Let $\phi_k^{h_n}$ denote the unique Kantorovich potential from $\mu_{kh_n}^{h_n}$ to $\mu_{(k-1)h_n}^{h_n}$. Using [15, Propositions 1.5.7 and 5.1.7], as well as [26, Equation (38)] with $\Psi = 0$, and optimality of $\mu_{kh_n}^{h_n}$, we get:

$$\frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\phi_k^{h_n}(x), \nabla\xi_k^n(x) \rangle d\mu_{kh_n}^{h_n}(x) - \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} (\psi_{kh_n}^{h_n})'(\theta^*) \langle \theta, \nabla\xi_k^n(x) \rangle d\theta d\mu_{kh_n}^{h_n}(x) \\ - \lambda \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi_k^n(x) d\mu_{kh_n}^{h_n}(x), \tag{25}$$

which is the derivative of $\mathcal{F}_\lambda^\nu(\cdot) + \frac{1}{2h_n}\mathcal{W}_2^2(\cdot, \mu_{(k-1)h_n})$ in the direction given by vector field $\nabla\xi_k^n$ is zero.

Let γ be the optimal transport between $\mu_{kh_n}^{h_n}$ and $\mu_{(k-1)h_n}^{h_n}$. Then:

$$\int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n}(x) - \rho_{(k-1)h_n}^{h_n}(x)}{h_n} dx = \frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} (\xi_k^n(y) - \xi_k^n(x)) d\gamma(x,y). \tag{26}$$

$$\frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\phi_k^{h_n}(x), \nabla\xi_k^n(x) \rangle d\mu_{kh_n}^{h_n}(x) = \frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\xi_k^n(x), y - x \rangle d\gamma(x,y). \tag{27}$$

Since ξ is C_c^∞, it has Lipschitz gradient. Let C be twice the Lipschitz constant of $\nabla\xi$. Then we have $|\xi(y) - \xi(x) - \langle \nabla\xi(x), y - x \rangle| \leq C|x-y|^2$, and hence:

$$\int_{\overline{\mathrm{B}}(0,r)} |\xi_k^n(y) - \xi_k^n(x) - \langle \nabla\xi_k^n(x), y - x \rangle| d\gamma(x,y) \leq C\mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}). \tag{28}$$

Combining (26), (27) and (28), we get:

$$\left|\sum_{k=1}^{\infty} h_n \int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n} - \rho_{(k-1)h_n}^{h_n}}{h_n} dx + \sum_{k=1}^{\infty} h_n \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla \phi_k^{h_n}, \nabla \xi_k^n \rangle d\mu_{kh_n}^{h_n}\right| \leq C \sum_{k=1}^{\infty} \mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}). \quad (29)$$

As some $\mathcal{F}_\lambda^\nu$ have a finite minimum on $\mathcal{P}(\overline{\mathrm{B}}(0,r))$, we have:

$$\begin{aligned} \sum_{k=1}^{\infty} \mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}) &\leq 2h_n \sum_{k=1}^{\infty} \mathcal{F}_\lambda^\nu(\mu_{(k-1)h_n}^{h_n}) - \mathcal{F}_\lambda^\nu(\mu_{kh_n}^{h_n}) \\ &\leq 2h_n \left(\mathcal{F}_\lambda^\nu(\mu_0) - \min_{\mathcal{P}(\overline{\mathrm{B}}(0,r))} \mathcal{F}_\lambda^\nu \right). \end{aligned} \quad (30)$$

and so the sum on the right hand side of the equation goes to zero as n goes to infinity.

From (29), (30) and (25) we conclude:

$$\lim_{n\to\infty} -h_n \sum_{k=1}^{\infty} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n} - \rho_{(k-1)h_n}^{h_n}}{h_n} dx = \lim_{n\to\infty} \left(h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} \psi_{kh_n,\theta}^{h_n}(\theta^*) \langle \theta, \nabla \xi_k^n \rangle d\theta d\mu_{kh_n}^{h_n} + h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta \xi_k^n(x) \rho_{kh_n}^{h_n}(x) dx \right), \quad (31)$$

where both limits exist, since the difference of left hand side and right hand side of the equation goes to zero, while the left hand side converges to a finite value by (22).

(5) Combining (22), (23), (24) and (31) we get the result.

□

7 Proof of Theorem 3

Before proceeding to the proof, let us first define the following Euler-Maruyama scheme which will be useful for our analysis:

$$\hat{X}_{k+1} = \hat{X}_k + h\hat{v}(\hat{X}_k, \mu_{kh}) + \sqrt{2\lambda h} Z_{n+1}, \quad (32)$$

where μ_t denotes the probability distribution of X_t with $(X_t)_t$ being the solution of the original SDE (7). Now, consider the probability distribution of $\hat{X}_k$ as $\hat{\mu}_{kh}$. Starting from the discrete-time process $(\hat{X}_k)_{k\in\mathbb{N}_+}$, we first define a continuous-time process $(Y_t)_{t\geq 0}$ that linearly interpolates $(\hat{X}_k)_{k\in\mathbb{N}_+}$, given as follows:

$$dY_t = \tilde{v}_t(Y)dt + \sqrt{2\lambda}dW_t, \tag{33}$$

where $\tilde{v}_t(Y) \triangleq -\sum_{k=0}^{\infty} \hat{v}_{kh}(Y_{kh})\mathbb{1}_{[kh,(k+1)h)}(t)$ and $\mathbb{1}$ denotes the indicator function. Similarly, we define a continuous-time process $(U_t)_{t\geq 0}$ that linearly interpolates $(\bar{X}_k)_{k\in\mathbb{N}_+}$, defined by (11), given as follows:

$$dU_t = \bar{v}_t(U)dt + \sqrt{2\lambda}dW_t, \tag{34}$$

where $\bar{v}_t(U) \triangleq -\sum_{k=0}^{\infty} \hat{v}(U_{kh}, \bar{\mu}_{kh})\mathbb{1}_{[kh,(k+1)h)}(t)$ and $\bar{\mu}_{kh}$ denotes the probability distribution of $\bar{X}_k$. Let us denote the distributions of $(X_t)_{t\in[0,T]}$, $(Y_t)_{t\in[0,T]}$ and $(U_t)_{t\in[0,T]}$ as π_X^T, π_Y^T and π_U^T respectively with $T = Kh$.

We consider the following assumptions:

H 1. *For all $\lambda > 0$, the SDE (7) has a unique strong solution denoted by $(X_t)_{t\geq 0}$ for any starting point $x \in \mathbb{R}^d$.*

H 2. *There exits $L < \infty$ such that*

$$\|v_t(x) - v_{t'}(x')\| \leq L(\|x - x'\| + |t - t'|), \tag{35}$$

where $v_t(x) = v(x, \mu_t)$ and

$$\|\hat{v}(x, \mu) - \hat{v}(x', \mu')\| \leq L(\|x - x'\| + \|\mu - \mu'\|_{\mathrm{TV}}). \tag{36}$$

H 3. *For all $t \geq 0$, v_t is dissipative, i.e. for all $x \in \mathbb{R}^d$,*

$$\langle x, v_t(x)\rangle \geq m\|x\|^2 - b, \tag{37}$$

for some $m, b > 0$.

H 4. *The estimator of the drift satisfies the following conditions: $\mathbb{E}[\hat{v}_t] = v_t$ for all $t \geq 0$, and for all $t \geq 0$, $x \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\hat{v}(x, \mu_t) - v(x, \mu_t)\|^2] \leq 2\delta(L^2\|x\|^2 + B^2), \tag{38}$$

for some $\delta \in (0, 1)$.

H 5. *For all $t \geq 0$: $|\Psi_t(0)| \leq A$ and $\|v_t(0)\| \leq B$, for $A, B \geq 0$, where $\Psi_t = \int_{\mathbb{S}^{d-1}} \psi_t(\langle\theta, \cdot\rangle)d\theta$.*

We start by upper-bounding $\|\hat{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}$.

Lemma 1. *Assume that the conditions **H**2 to 5 hold. Then, the following bound holds:*

$$\|\hat{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}^2 \leq \|\pi_Y^T - \pi_X^T\|_{\mathrm{TV}}^2 \leq \frac{L^2 K}{4\lambda}\Big(\frac{C_1 h^3}{3} + 3\lambda d h^2\Big) + \frac{C_2 \delta K h}{8\lambda}, \tag{39}$$

where $C_1 \triangleq 12(L^2 C_0 + B^2) + 1$, $C_2 \triangleq 2(L^2 C_0 + B^2)$, $C_0 \triangleq C_e + 2(1 \vee \frac{1}{m})(b + 2B^2 + d\lambda)$, *and* C_e *denotes the entropy of* μ_0.

Proof. We use the proof technique presented in [17, 38]. It is easy to verify that for all $k \in \mathbb{N}_+$, we have $Y_{kh} = \hat{X}_k$.

By Girsanov's theorem to express the Kullback-Leibler (KL) divergence between these two distributions, given as follows:

$$\begin{aligned}
\mathrm{KL}(\pi_X^T || \pi_Y^T) &= \frac{1}{4\lambda}\int_0^{Kh} \mathbb{E}[\|v_t(Y_t) + \tilde{v}_t(Y)\|^2]\, dt && (40)\\
&= \frac{1}{4\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) + \tilde{v}_t(Y)\|^2]\, dt && (41)\\
&= \frac{1}{4\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt. && (42)
\end{aligned}$$

By using $v_t(Y_t) - \hat{v}_{kh}(Y_{kh}) = (v_t(Y_t) - v_{kh}(Y_{kh})) + (v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh}))$, we obtain

$$\begin{aligned}
\mathrm{KL}(\pi_X^T || \pi_Y^T) \leq& \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) - v_{kh}(Y_{kh})\|^2]\, dt\\
&+ \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt && (43)\\
\leq& \frac{L^2}{\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \big(\mathbb{E}[\|Y_t - Y_{kh}\|^2] + (t-kh)^2\big)\, dt\\
&+ \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt. && (44)
\end{aligned}$$

The last inequality is due to the Lipschitz condition **H**2.

Now, let us focus on the term $\mathbb{E}[\|Y_t - Y_{kh}\|^2]$. By using (33), we obtain:

$$Y_t - Y_{kh} = -(t-kh)\hat{v}_{kh}(Y_{kh}) + \sqrt{2\lambda(t-kh)}Z, \tag{45}$$

where Z denotes a standard normal random variable. By adding and subtracting the term $-(t-kh)v_{kh}(Y_{kh})$, we have:

$$Y_t - Y_{kh} = -(t-kh)v_{kh}(Y_{kh}) + (t-kh)(v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})) + \sqrt{2\lambda(t-kh)}Z. \tag{46}$$

Taking the square and then the expectation of both sides yields:

$$\begin{aligned}\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq & 3(t-kh)^2\mathbb{E}[\|v_{kh}(Y_{kh})\|^2] + 3(t-kh)^2\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2] \\ & + 6\lambda(t-kh)d. \end{aligned} \tag{47}$$

As a consequence of **H** 2 and **H** 5, we have $\|v_t(x)\| \leq L\|x\| + B$ for all $t \geq 0$, $x \in \mathbb{R}^d$. Combining this inequality with **H**4, we obtain:

$$\begin{aligned}\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq & 6(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) + 6(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) \\ & + 6\lambda(t-kh)d \qquad (48) \\ = & 12(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) + 6\lambda(t-kh)d. \qquad (49)\end{aligned}$$

By Lemma 3.2 of [17][3], we have $\mathbb{E}[\|Y_{kh}\|^2] \leq C_0 \triangleq C_e + 2(1 \vee \frac{1}{m})(b + 2B^2 + d\lambda)$, where C_e denotes the entropy of μ_0. Using this result in the above equation yields:

$$\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq 12(t-kh)^2(L^2C_0 + B^2) + 6\lambda(t-kh)d. \tag{50}$$

We now focus on the term $\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]$ in (44). Similarly to the previous term, we can upper-bound this term as follows:

$$\begin{aligned}\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2] \leq & 2\delta(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) \qquad (51) \\ \leq & 2\delta(L^2C_0 + B^2). \qquad (52)\end{aligned}$$

By using (50) and (52) in (44), we obtain:

$$\begin{aligned}\mathrm{KL}(\pi_X^T||\pi_Y^T) \leq & \frac{L^2}{\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\big(12(t-kh)^2(L^2C_0 + B^2) + 6\lambda(t-kh)d + (t-kh)^2\big)dt \\ & + \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} 2\delta(L^2C_0 + B^2)\, dt \qquad (53) \\ = & \frac{L^2K}{\lambda}\Big(\frac{C_1h^3}{3} + \frac{6\lambda dh^2}{2}\Big) + \frac{C_2\delta Kh}{2\lambda}, \qquad (54)\end{aligned}$$

where $C_1 = 12(L^2C_0 + B^2) + 1$ and $C_2 = 2(L^2C_0 + B^2)$.

Finally, by using the data processing and Pinsker inequalities, we obtain:

$$\begin{aligned}\|\hat{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}^2 \leq \|\pi_X^T - \pi_Y^T\|_{\mathrm{TV}}^2 \leq & \frac{1}{4}\mathrm{KL}(\pi_X^T||\pi_Y^T) \qquad (55) \\ = & \frac{L^2K}{4\lambda}\Big(\frac{C_1h^3}{3} + 3\lambda dh^2\Big) + \frac{C_2\delta Kh}{8\lambda}. \qquad (56)\end{aligned}$$

This concludes the proof. □

[3]Note that Lemma 3.2 of [17] considers the case where the drift is not time- or measure-dependent. However, with **H**3 it is easy to show that the same result holds for our case as well.

Now, we bound the term $\|\bar{\mu}_{Kh} - \hat{\mu}_{Kh}\|_{\mathrm{TV}}$.

Lemma 2. *Assume that **H**2 holds. Then the following bound holds:*

$$\|\pi_U^T - \pi_Y^T\|_{\mathrm{TV}}^2 \leq \frac{L^2 Kh}{16\lambda} \|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2. \tag{57}$$

Proof. We use that same approach than in Lemma 1. By Girsanov's theorem once again, we have

$$\mathrm{KL}(\pi_Y^T || \pi_U^T) = \frac{1}{4\lambda} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}[\|\hat{v}(U_{kh}, \mu_{kh}) - \hat{v}(U_{kh}, \bar{\mu}_{kh})\|^2] \, dt, \tag{58}$$

where π_U^T denotes the distributions of $(U_t)_{t\in[0,T]}$ with $T = Kh$. By using $\mathbf{H}2$, we have:

$$\mathrm{KL}(\pi_Y^T || \pi_U^T) \leq \frac{L^2 h}{4\lambda} \sum_{k=0}^{K-1} \|\mu_{kh} - \bar{\mu}_{kh}\|_{\mathrm{TV}}^2 \tag{59}$$

$$\leq \frac{L^2 Kh}{4\lambda} \|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2. \tag{60}$$

By applying the data processing and Pinsker inequalities, we obtain the desired result. □

7.1 Proof of Theorem 3

Here, we precise the statement of Theorem 3.

Theorem 7. *Assume that the assumptions in Lemma 1 and Lemma 2 hold. Then for $\lambda > \frac{KL^2h}{8}$, the following bound holds:*

$$\|\bar{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}^2 \leq \delta_\lambda \left\{ \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} \right\}, \tag{61}$$

where $\delta_\lambda = (1 - \frac{KL^2h}{8\lambda})^{-1}$.

Proof. We have the following decomposition: (with $T = Kh$)

$$\|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2 \leq 2\|\pi_X^T - \pi_Y^T\|_{\mathrm{TV}}^2 + 2\|\pi_Y^T - \pi_U^T\|_{\mathrm{TV}}^2 \tag{62}$$

$$\leq \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} + \frac{L^2 K h}{8\lambda} \|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2 \tag{63}$$

$$\leq \Big(1 - \frac{KL^2 h}{8\lambda} \Big)^{-1} \left\{ \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} \right\}. \tag{64}$$

The second line follows from Lemma 1 and Lemma 2. Last line follows from the assumption that λ is large enough. This completes the proof. □

8 Proof of Corollary 1

Proof. Considering the bound given in Theorem 3, the choice h implies that

$$\frac{\delta_\lambda L^2 K}{2\lambda}\Big(\frac{C_1 h^3}{3} + 3\lambda d h^2\Big) \leq \varepsilon^2. \tag{65}$$

This finalizes the proof. □