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

SUPPLEMENTARY DOCUMENT

Antoine Liutkus [1] **Umut Şimşekli** [2] **Szymon Majewski** [3] **Alain Durmus** [4] **Fabian-Robert Stöter** [1]

1. Proof of Theorem 2

We first need to generalize (Bonnotte, 2013)[Lemma 5.4.3] to distribution $\rho \in \mathrm{L}^\infty(\overline{\mathrm{B}}(0,r))$, $r > 0$.

Theorem S1. *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:*

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

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

$$||\rho||_{\mathrm{L}^\infty} \leq (1 + h/\sqrt{d})^d ||\rho_0||_{\mathrm{L}^\infty}. \tag{S1}$$

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 (Ambrosio et al., 2008)[Lemma 9.4.3] $\mathcal{H}$ is lower semicontinuous on $(\mathcal{P}(\overline{\mathrm{B}}(0,r)), \mathcal{W}_2)$. Since by (Bonnotte, 2013)[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 S1 is true by (Bonnotte, 2013)[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 $\mathcal{F}_\lambda^\nu(\cdot) + \frac{1}{2h}\mathcal{W}_2^2(\cdot,\mu_t)$, and let $\hat{\rho}_t$ be the density of $\hat{\mu}_t$. Using (Bonnotte, 2013)[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

[1]Inria and LIRMM, Univ. of Montpellier, France [2]LTCI, Télécom Paristech, Université Paris-Saclay, Paris, France [3]Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland [4]CNRS, ENS Paris-Saclay,Université Paris-Saclay, Cachan, France. Correspondence to: Antoine Liutkus <antoine.liutkus@inria.fr>, Umut Simsekli <umut.simsekli@telecom-paristech.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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 S1. □

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 S2. *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*

$$\mathcal{F}^\nu(h,\mu_+,\mu_-) = \mathcal{F}^\nu_\lambda(\mu_+) + \frac{1}{2h}\mathcal{W}_2^2(\mu_+,\mu_-). \tag{S2}$$

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

$$||\rho_t||_{\mathrm{L}^\infty} \leq e^{dt\sqrt{d}}||\rho_0||_{\mathrm{L}^\infty}. \tag{S3}$$

Proof. We start by noting, that by S1 for any $h>0$ there exists a piecewise constant trajectory μ^h for S2 starting at μ_0. Furthermore for $t\geq0$ measure $\mu^h_t=\mu^h(t)$ has density ρ^h_t, and:

$$||\rho^h_t||_{\mathrm{L}^\infty} \leq e^{d\sqrt{d}(t+h)}||\rho_0||_{\mathrm{L}^\infty}. \tag{S4}$$

Let us choose $T>0$. We denote $\rho^h(t,x)=\rho^h_t(x)$. For $h\leq1$, 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:

$$\begin{aligned}\left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_n}-\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_m}\right|\leq\\ \left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_n}-\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_n}ds\right|+\left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_m}-\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_m}ds\right|+\\ \left|\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_m}ds-\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_n}ds\right|.\end{aligned}\tag{S5}$$

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:

$$\left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_n}-\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_n}\right|^2\quad\leq\quad\int_{\overline{\mathrm{B}}(0,r)\times\overline{\mathrm{B}}(0,r)}|\zeta(x)-\zeta(y)|^2\,d\gamma(x,y)\quad\leq\quad||\nabla\zeta||_\infty^2\mathcal{W}_2^2(\mu_t^{h_n},\mu_s^{h_n}).\tag{S6}$$

In addition, for $n_t=\lfloor t/h_n\rfloor$ and $n_s=\lfloor s/h_n\rfloor$ we have $\mu_t^{h_n}=\mu_{n_th_n}^{h_n}$ and $\mu_s^{h_n}=\mu_{n_sh_n}^{h_n}$. For all $k\geq0$ we have:

$$\mathcal{W}_2^2(\mu_{kh_n}^{h_n},\mu_{(k+1)h_n}^{h_n})\leq2h_n(\mathcal{F}_\lambda^\nu(\mu_{kh_n}^{h_n})-\mathcal{F}_\lambda^\nu(\mu_{(k+1)h_n}^{h_n}).\tag{S7}$$

Using this result and (S6) and assuming without loss of generality $n_t\leq n_s$, from the Cauchy-Schwartz inequality we get:

$$\begin{aligned}\mathcal{W}_2^2(\mu_t^{h_n},\mu_s^{h_n})&\leq\left(\sum_{k=n_t}^{n_s-1}\mathcal{W}_2(\mu_{kh_n}^{h_n},\mu_{(k+1)h_n}^{h_n})\right)^2\\&\leq|n_t-n_s|\sum_{k=n_t}^{n_s1}\mathcal{W}_2^2(\mu_{kh_n}^{h_n},\mu_{(k+1)h_n}^{h_n})\\&\leq2h_n|n_t-n_s|(\mathcal{F}_\lambda^\nu(\mu_{n_th_n}^{h_n})-\mathcal{F}_\lambda^\nu(\mu_{n_sh_n}^{h_n}))\leq2C(|t-s|+h_n),\end{aligned}\tag{S8}$$

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 (S7) 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 S6 we get:

$$\begin{aligned}\left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_n}-\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_n}ds\right|^2&\leq\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}\left|\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_t^{h_n}-\int_{\overline{\mathrm{B}}(0,r)}\zeta d\mu_s^{h_n}\right|^2ds\\&\leq\frac{C||\nabla\zeta||_\infty^2}{\delta}\int_{t-\delta}^{t+\delta}(|t-s|+h_n)ds\\&\leq2C||\nabla\zeta||_\infty^2(h_n+\delta).\end{aligned}$$

Together with (S5), 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 (S8), 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 (S2) follows from (S4). □

Theorem S3. *Let $(\mu_t)_{t\geq 0}$ be a generalized minimizing movement scheme given by Theorem S2 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:

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

Proof. Our proof is based on the proof of (Bonnotte, 2013)[Theorem 5.6.1]. We proceed in five steps.

(1) Let $h_n \to 0$ be a sequence given by Theorem S2, 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 (Bonnotte, 2013)[Theorem 5.6.1], we obtain:

$$\int_{\overline{\mathrm{B}}(0,r)} \xi(0,x)\rho_0(x)dx + \int_0^\infty \int_{\overline{\mathrm{B}}(0,r)} \frac{\partial \xi}{\partial t}(t,x)\rho_t(x)dxdt$$
$$= \lim_{n\to\infty} -h_n \sum_{k=1}^\infty \int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n}(x) - \rho_{(k-1)h_n}^{h_n}(x)}{h_n} dx. \quad \text{(S9)}$$

(2) Again, this part is the same as part 2 of the proof of (Bonnotte, 2013)[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 (Bonnotte, 2013)[Theorem 5.6.1], we get:

$$\int_0^\infty \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} (\psi_{t,\theta})'(\langle\theta,x\rangle)\langle\theta, \nabla\xi(x,t)\rangle d\theta d\mu_t(x)dt$$
$$= \lim_{n\to\infty} h_n \sum_{k=1}^\infty \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} \psi_{kh_n,\theta}^{h_n}(\theta^*)\langle\theta, \nabla\xi_k^n\rangle d\theta d\mu_{kh_n}^{h_n}. \quad \text{(S10)}$$

(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:

$$\lim_{n \to \infty} h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi_k^n(x) \rho_{kh_n}^{h_n}(x) dx = \int_0^{+\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi(t,x) \rho_t(x) dx dt. \tag{S11}$$

(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 (Bonnotte, 2013)[Propositions 1.5.7 and 5.1.7], as well as (Jordan et al., 1998)[Equation (38)] with $\Psi = 0$, and optimality of $\mu_{kh_n}^{h_n}$, we get:

$$\begin{aligned} \frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\phi_k^{h_n}(x), \nabla\xi_k^n(x) \rangle d\mu_{kh_n}^{h_n}(x) - \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} (\psi_{kh_n}^{h_n})'(\theta^*) \langle \theta, \nabla\xi_k^n(x) \rangle d\theta d\mu_{kh_n}^{h_n}(x) \\ - \lambda \int_{\overline{\mathrm{B}}(0,r)} \Delta\xi_k^n(x) d\mu_{kh_n}^{h_n}(x), \end{aligned} \tag{S12}$$

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:

$$\int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n}(x) - \rho_{(k-1)h_n}^{h_n}(x)}{h_n} dx = \frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} (\xi_k^n(y) - \xi_k^n(x)) d\gamma(x,y). \tag{S13}$$

$$\frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\phi_k^{h_n}(x), \nabla\xi_k^n(x) \rangle d\mu_{kh_n}^{h_n}(x) = \frac{1}{h_n} \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\xi_k^n(x), y - x \rangle d\gamma(x,y). \tag{S14}$$

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:

$$\int_{\overline{\mathrm{B}}(0,r)} |\xi_k^n(y) - \xi_k^n(x) - \langle \nabla\xi_k^n(x), y - x \rangle| d\gamma(x,y) \leq C\mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}). \tag{S15}$$

Combining (S13), (S14) and (S15), we get:

$$\begin{aligned} \left| \sum_{k=1}^{\infty} h_n \int_{\overline{\mathrm{B}}(0,r)} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n} - \rho_{(k-1)h_n}^{h_n}}{h_n} dx + \sum_{k=1}^{\infty} h_n \int_{\overline{\mathrm{B}}(0,r)} \langle \nabla\phi_k^{h_n}, \nabla\xi_k^n \rangle d\mu_{kh_n}^{h_n} \right| \\ \leq C \sum_{k=1}^{\infty} \mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}). \end{aligned} \tag{S16}$$

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

$$\begin{aligned} \sum_{k=1}^{\infty} \mathcal{W}_2^2(\mu_{(k-1)h_n}^{h_n}, \mu_{kh_n}^{h_n}) &\leq 2h_n \sum_{k=1}^{\infty} \mathcal{F}_\lambda^\nu(\mu_{(k-1)h_n}^{h_n}) - \mathcal{F}_\lambda^\nu(\mu_{kh_n}^{h_n}) \\ &\leq 2h_n \left(\mathcal{F}_\lambda^\nu(\mu_0) - \min_{\mathcal{P}(\overline{\mathrm{B}}(0,r))} \mathcal{F}_\lambda^\nu \right). \end{aligned} \tag{S17}$$

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

From (S16), (S17) and (S12) we conclude:

$$\lim_{n\to\infty} -h_n \sum_{k=1}^{\infty} \xi_k^n(x) \frac{\rho_{kh_n}^{h_n} - \rho_{(k-1)h_n}^{h_n}}{h_n} dx = \lim_{n\to\infty} \left(h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \int_{\mathbb{S}^{d-1}} \psi_{kh_n,\theta}^{h_n}(\theta^*) \langle \theta, \nabla \xi_k^n \rangle d\theta d\mu_{kh_n}^{h_n} + h_n \sum_{k=1}^{\infty} \int_{\overline{\mathrm{B}}(0,r)} \Delta \xi_k^n(x) \rho_{kh_n}^{h_n}(x) dx \right), \quad \text{(S18)}$$

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 (S9).

(5) Combining (S9), (S10), (S11) and (S18) we get the result.

□

2. 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:

$$\hat{X}_{k+1} = \hat{X}_k + h\hat{v}(\hat{X}_k, \mu_{kh}) + \sqrt{2\lambda h} Z_{n+1}, \quad \text{(S19)}$$

where μ_t denotes the probability distribution of X_t with $(X_t)_t$ being the solution of the original SDE (8). 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:

$$dY_t = \tilde{v}_t(Y)dt + \sqrt{2\lambda} dW_t, \quad \text{(S20)}$$

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 (13), given as follows:

$$dU_t = \bar{v}_t(U)dt + \sqrt{2\lambda} dW_t, \quad \text{(S21)}$$

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:

HS1. *For all $\lambda > 0$, the SDE (8) has a unique strong solution denoted by $(X_t)_{t\geq 0}$ for any starting point $x \in \mathbb{R}^d$.*

HS2. *There exits $L < \infty$ such that*

$$\|v_t(x) - v_{t'}(x')\| \leq L(\|x - x'\| + |t - t'|), \quad \text{(S22)}$$

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

$$\|\hat{v}(x,\mu) - \hat{v}(x',\mu')\| \leq L(\|x - x'\| + \|\mu - \mu'\|_{\mathrm{TV}}). \quad \text{(S23)}$$

HS3. *For all $t \geq 0$, v_t is dissipative, i.e. for all $x \in \mathbb{R}^d$,*

$$\langle x, v_t(x) \rangle \geq m\|x\|^2 - b, \quad \text{(S24)}$$

for some $m, b > 0$.

HS4. *The estimator of the drift satisfies the following conditions:* $\mathbb{E}[\hat{v}_t] = v_t$ *for all* $t \geq 0$*, and for all* $t \geq 0$*,* $x \in \mathbb{R}^d$*,*

$$\mathbb{E}[\|\hat{v}(x, \mu_t) - v(x, \mu_t)\|^2] \leq 2\delta(L^2\|x\|^2 + B^2), \tag{S25}$$

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

HS5. *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 $\|\hat{\mu}_{Kh} - \mu_T\|_{\text{TV}}$.

Lemma S1. *Assume that the conditions* ***H****S2 to S5 hold. Then, the following bound holds:*

$$\|\hat{\mu}_{Kh} - \mu_T\|_{\text{TV}}^2 \leq \|\pi_Y^T - \pi_X^T\|_{\text{TV}}^2 \leq \frac{L^2K}{4\lambda}\Big(\frac{C_1h^3}{3} + 3\lambda dh^2\Big) + \frac{C_2\delta Kh}{8\lambda}, \tag{S26}$$

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

Proof. We use the proof technique presented in (Dalalyan, 2017; Raginsky et al., 2017). 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:

$$\begin{aligned}
\text{KL}(\pi_X^T||\pi_Y^T) &= \frac{1}{4\lambda}\int_0^{Kh} \mathbb{E}[\|v_t(Y_t) + \tilde{v}_t(Y)\|^2]\, dt && \text{(S27)}\\
&= \frac{1}{4\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) + \tilde{v}_t(Y)\|^2]\, dt && \text{(S28)}\\
&= \frac{1}{4\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt. && \text{(S29)}
\end{aligned}$$

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

$$\begin{aligned}
\text{KL}(\pi_X^T||\pi_Y^T) \leq& \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_t(Y_t) - v_{kh}(Y_{kh})\|^2]\, dt\\
&+ \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt && \text{(S30)}\\
\leq& \frac{L^2}{\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \big(\mathbb{E}[\|Y_t - Y_{kh}\|^2] + (t - kh)^2\big)\, dt\\
&+ \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]\, dt. && \text{(S31)}
\end{aligned}$$

The last inequality is due to the Lipschitz condition **H**S2.

Now, let us focus on the term $\mathbb{E}[\|Y_t - Y_{kh}\|^2]$. By using (S20), we obtain:

$$Y_t - Y_{kh} = -(t - kh)\hat{v}_{kh}(Y_{kh}) + \sqrt{2\lambda(t - kh)}Z, \tag{S32}$$

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

$$Y_t - Y_{kh} = -(t - kh)v_{kh}(Y_{kh}) + (t - kh)(v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})) + \sqrt{2\lambda(t - kh)}Z. \tag{S33}$$

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

$$\begin{aligned}
\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq& 3(t - kh)^2\mathbb{E}[\|v_{kh}(Y_{kh})\|^2] + 3(t - kh)^2\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]\\
&+ 6\lambda(t - kh)d. && \text{(S34)}
\end{aligned}$$

As a consequence of **H**S2 and **H**S5, we have $\|v_t(x)\| \leq L\|x\| + B$ for all $t \geq 0$, $x \in \mathbb{R}^d$. Combining this inequality with **H** S4, we obtain:

$$\begin{aligned}\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq & 6(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) + 6(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) \\ & + 6\lambda(t-kh)d \qquad \text{(S35)} \\ = & 12(t-kh)^2(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) + 6\lambda(t-kh)d. \qquad \text{(S36)}\end{aligned}$$

By Lemma 3.2 of (Raginsky et al., 2017)[1], 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:

$$\mathbb{E}[\|Y_t - Y_{kh}\|^2] \leq 12(t-kh)^2(L^2C_0 + B^2) + 6\lambda(t-kh)d. \qquad \text{(S37)}$$

We now focus on the term $\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2]$ in (S31). Similarly to the previous term, we can upper-bound this term as follows:

$$\begin{aligned}\mathbb{E}[\|v_{kh}(Y_{kh}) - \hat{v}_{kh}(Y_{kh})\|^2] \leq & 2\delta(L^2\mathbb{E}[\|Y_{kh}\|^2] + B^2) \qquad \text{(S38)} \\ \leq & 2\delta(L^2C_0 + B^2). \qquad \text{(S39)}\end{aligned}$$

By using (S37) and (S39) in (S31), we obtain:

$$\begin{aligned}\mathrm{KL}(\pi_X^T||\pi_Y^T) \leq & \frac{L^2}{\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\big(12(t-kh)^2(L^2C_0 + B^2) + 6\lambda(t-kh)d + (t-kh)^2\big)dt \\ & + \frac{1}{2\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} 2\delta(L^2C_0 + B^2)\,dt \qquad \text{(S40)} \\ = & \frac{L^2K}{\lambda}\Big(\frac{C_1h^3}{3} + \frac{6\lambda dh^2}{2}\Big) + \frac{C_2\delta Kh}{2\lambda}, \qquad \text{(S41)}\end{aligned}$$

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:

$$\begin{aligned}\|\hat{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}^2 \leq \|\pi_X^T - \pi_Y^T\|_{\mathrm{TV}}^2 \leq & \frac{1}{4}\mathrm{KL}(\pi_X^T||\pi_Y^T) \qquad \text{(S42)} \\ = & \frac{L^2K}{4\lambda}\Big(\frac{C_1h^3}{3} + 3\lambda dh^2\Big) + \frac{C_2\delta Kh}{8\lambda}. \qquad \text{(S43)}\end{aligned}$$

This concludes the proof. □

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

Lemma S2. *Assume that **H**S2 holds. Then the following bound holds:*

$$\|\pi_U^T - \pi_Y^T\|_{\mathrm{TV}}^2 \leq \frac{L^2Kh}{16\lambda}\|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2. \qquad \text{(S44)}$$

Proof. We use that same approach than in Lemma S1. By Girsanov's theorem once again, we have

$$\mathrm{KL}(\pi_Y^T||\pi_U^T) = \frac{1}{4\lambda}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\mathbb{E}[\|\hat{v}(U_{kh}, \mu_{kh}) - \hat{v}(U_{kh}, \bar{\mu}_{kh})\|^2]\,dt, \qquad \text{(S45)}$$

[1]Note that Lemma 3.2 of (Raginsky et al., 2017) considers the case where the drift is not time- or measure-dependent. However, with **H**S3 it is easy to show that the same result holds for our case as well.

where π_U^T denotes the distributions of $(U_t)_{t\in[0,T]}$ with $T = Kh$. By using **H**S2, we have:

$$\mathrm{KL}(\pi_Y^T||\pi_U^T) \leq \frac{L^2 h}{4\lambda} \sum_{k=0}^{K-1} \|\mu_{kh} - \bar{\mu}_{kh}\|_{\mathrm{TV}}^2 \tag{S46}$$

$$\leq \frac{L^2 Kh}{4\lambda} \|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2. \tag{S47}$$

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

2.1. Proof of Theorem 3

Here, we precise the statement of Theorem 3.

Theorem S4. *Assume that the assumptions in Lemma S1 and Lemma S2 hold. Then for $\lambda > \frac{KL^2h}{8}$, the following bound holds:*

$$\|\bar{\mu}_{Kh} - \mu_T\|_{\mathrm{TV}}^2 \leq \delta_\lambda \left\{ \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} \right\}, \tag{S48}$$

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

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

$$\|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2 \leq 2\|\pi_X^T - \pi_Y^T\|_{\mathrm{TV}}^2 + 2\|\pi_Y^T - \pi_U^T\|_{\mathrm{TV}}^2 \tag{S49}$$

$$\leq \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} + \frac{L^2 K h}{8\lambda} \|\pi_X^T - \pi_U^T\|_{\mathrm{TV}}^2 \tag{S50}$$

$$\leq \Big(1 - \frac{KL^2 h}{8\lambda} \Big)^{-1} \left\{ \frac{L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) + \frac{C_2 \delta K h}{4\lambda} \right\}. \tag{S51}$$

The second line follows from Lemma S1 and Lemma S2. Last line follows from the assumption that λ is large enough. This completes the proof. □

3. Proof of Corollary 1

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

$$\frac{\delta_\lambda L^2 K}{2\lambda} \Big(\frac{C_1 h^3}{3} + 3\lambda d h^2 \Big) \leq \varepsilon^2. \tag{S52}$$

This finalizes the proof. □

4. Additional Experimental Results

4.1. The Sliced Wasserstein Flow

The whole code for the Sliced Wasserstein Flow was implemented in Python, for use with Pytorch[2]. The code was written so as to run efficiently on GPU, and is available on the publicly available repository related to this paper[3].

In practice, the SWF involves relatively simple operations, the most important being:

- For each random $\theta \in \{\theta_n\}_{n=1\dots N_\theta}$, compute its inner product with all items from a dataset and obtain the empirical quantiles for these *projections*.

[2] http://www.pytorch.org.
[3] https://github.com/aliutkus/swf.

- At each step k of the SWF, for each projection $z = \langle \theta, \bar{X}_k^i \rangle$, apply two piece-wise linear functions, corresponding to the scalar optimal transport $\psi'_{k,\theta}(z)$.

Even if such steps are conceptually simple, the quantile and required linear interpolation functions were not available on GPU for any framework we could figure out at the time of writing this paper. Hence, we implemented them ourselves for use with Pytorch, and the interested reader will find the details in the Github repository dedicated to this paper.

Given these operations, putting a SWF implementation together is straightforward. The code provided allows not only to apply it on any dataset, but also provides routines to have the computation of these sketches running in the background in a parallel manner.

4.2. The need for dimension reduction through autoencoders

In this study, we used an autoencoder trained on the dataset as a dimension reduction technique, so that the SWF is applied to transport particles in a latent space of dimension $d \approx 50$, instead of the original $d > 1000$ of image data.

The curious reader may wonder why SWF is not applied directly to this original space, and what performances should be expected there. We have done this experiment, and we found out that SWF has much trouble rapidly converging to satisfying samples. In figure S1, we show the progressive evolution of particles undergoing SWF when the target is directly taken as the uncompressed dataset.

In this experiment, the strategy was to change the projections θ at each iteration, so that we ended up with a set of projections being $\{\theta_{n,k}\}_{n=1\dots N_\theta}^{k=1\dots K}$ instead of the fixed set of N_θ we now consider in the main document (for this, we picked $N_\theta = 200$). This strategy is motivated by the complete failure we observed whenever we picked such fixed projections throughout iterations, even for a relatively large number as $N_\theta = 16000$.

As may be seen on Figure S1, the particles definitely converge to samples from the desired datasets, and this is encouraging. However, we feel that the extreme number of iterations required to achieve such convergence comes from the fact that theory needs an integral over the $d-$dimensional sphere at each step of the SWF, which is clearly an issue whenever d gets too large. Although our solution of picking new samples from the sphere at each iteration alleviated this issue to some extent, the curse of dimensionality prevents us from doing much better with just thousands of *random* projections at a time.

This being said, we are confident that good performance would be obtained if millions of random projections could be considered for transporting such high dimensional data because i/ theory suggests it and ii/ we observed excellent performance on reduced dimensions.

However, we, unfortunately, did not have the computing power it takes for such large scale experiments and this is what motivated us in the first place to introduce some dimension-reduction technique through AE.

4.3. Structure of our autoencoders for reducing data dimension

As mentioned in the text, we used autoencoders to reduce the dimensionality of the transport problem. The structure of these networks is the following:

- **Encoder** Four 2d convolution layers with (num_chan_out, kernel_size, stride, padding) being $(3, 3, 1, 1)$, $(32, 2, 2, 0)$, $(32, 3, 1, 1)$, $(32, 3, 1, 1)$, each one followed by a ReLU activation. At the output, a linear layer gets the desired bottleneck size.

- **Decoder** A linear layer gets from the bottleneck features to a vector of dimension 8192, which is reshaped as $(32, 16, 16)$. Then, three convolution layers are applied, all with 32 output channels and (kernel_size, stride, panning) being respectively $(3, 1, 1)$, $(3, 1, 1)$, $(2, 2, 0)$. A 2d convolution layer is then applied with an output number of channels being that of the data (1 for black and white, 3 for color), and a (kernel_size, stride, panning) as $(3, 1, 1)$. In any case, all layers are followed by a ReLU activation, and a sigmoid activation is applied a the very output.

Once these networks defined, these autoencoders are trained in a very simple manner by minimizing the binary cross entropy between input and output over the training set of the considered dataset (here MNIST, CelebA or FashionMNIST). This training was achieved with the Adam algorithm (Kingma & Ba, 2014) with learning rate $1e-3$.

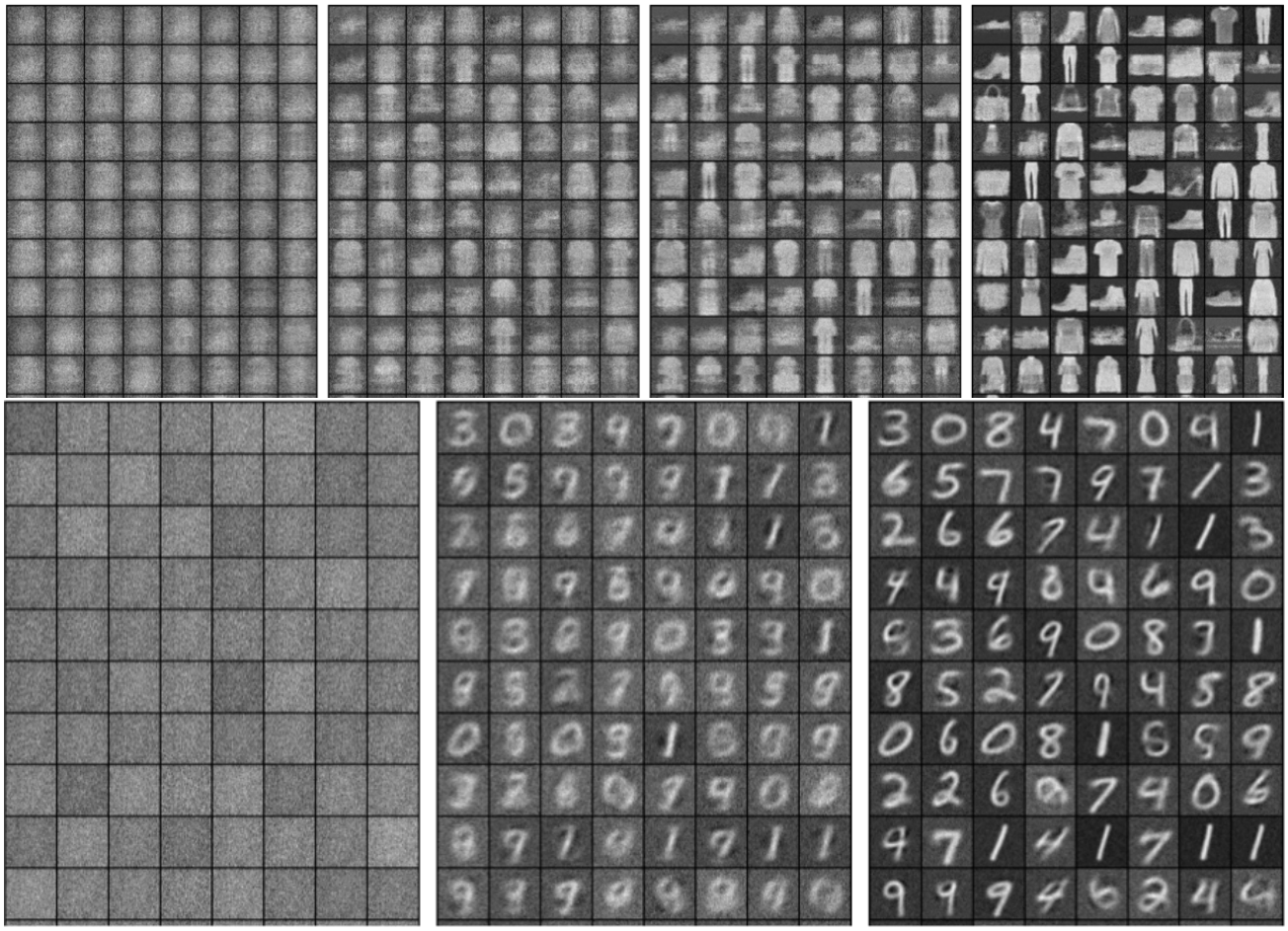


Figure S1. The evolution of SWF through 15000 iterations, when the original high-dimensional data is kept instead of working on reduced bottleneck features as done in the main document. Showing results on the MNIST and FashionMNIST datasets. For a visual comparison for FashionMNIST, we refer the reader to (Samangouei et al., 2018).

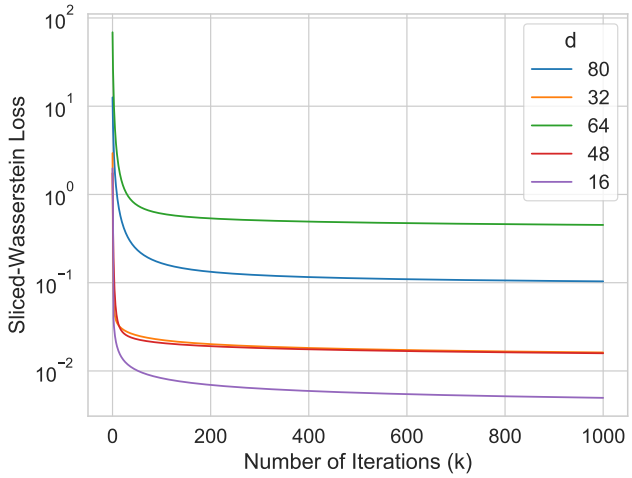


Figure S2. Approximately computed $\mathcal{SW}_2$ between the output $\bar{\mu}_k^N$ and data distribution ν in the MNIST experiment for different dimensions d for the bottleneck features (and the corresponding pre-trained AE).

No additional training trick was involved as in Variational Autoencoder (Kingma & Welling, 2013) to make sure the distribution of the bottleneck features matches some prior. The core advantage of the proposed method in this respect is indeed to turn any previously learned AE as a generative model, by automatically and non-parametrically transporting particles drawn from an arbitrary prior distribution μ to the observed empirical distribution ν of the bottleneck features over the training set.

4.4. Convergence plots of SWF

In the same experimental setting as in the main document, we also illustrate the behavior of the algorithm for varying dimensionality d for the bottleneck-features. To monitor the convergence of SWF as predicted by theory, we display 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 S2 illustrates the results. We observe that, for all choices of d, we see a steady and smooth decrease in the cost for all runs, which is in line with our theory. The absolute value of the cost for varying dimensions remains hard to interpret at this stage of our investigations.

5. Additional samples

5.1. Evolution throughout iterations

In Figures S3 and S4 below, we provide the evolution of the SWF algorithm on the Fashion MNIST and the MNIST datasets in higher resolution, for an AE with $d = 48$ bottleneck features.

5.2. Training samples, interpolation and extrapolation

In Figures S5 and S6 below, we provide other examples of outcome from SWF, both for the MNIST and the FashionMNIST datasets, still with $d = 48$ bottleneck features.

The most noticeable fact we may see on these figures is that while the actual particles which went through SWF, as well as linear combinations of them, all yield very satisfying results, this is however not the case for particles that are drawn randomly and then brought through a pre-learned SWF.

Once again, we interpret this fact through the curse of dimensionality: while we saw in our toy GMM example that using a pre-trained SWF was totally working for small dimensions, it is already not so for $d = 48$ and only 3000 training samples.

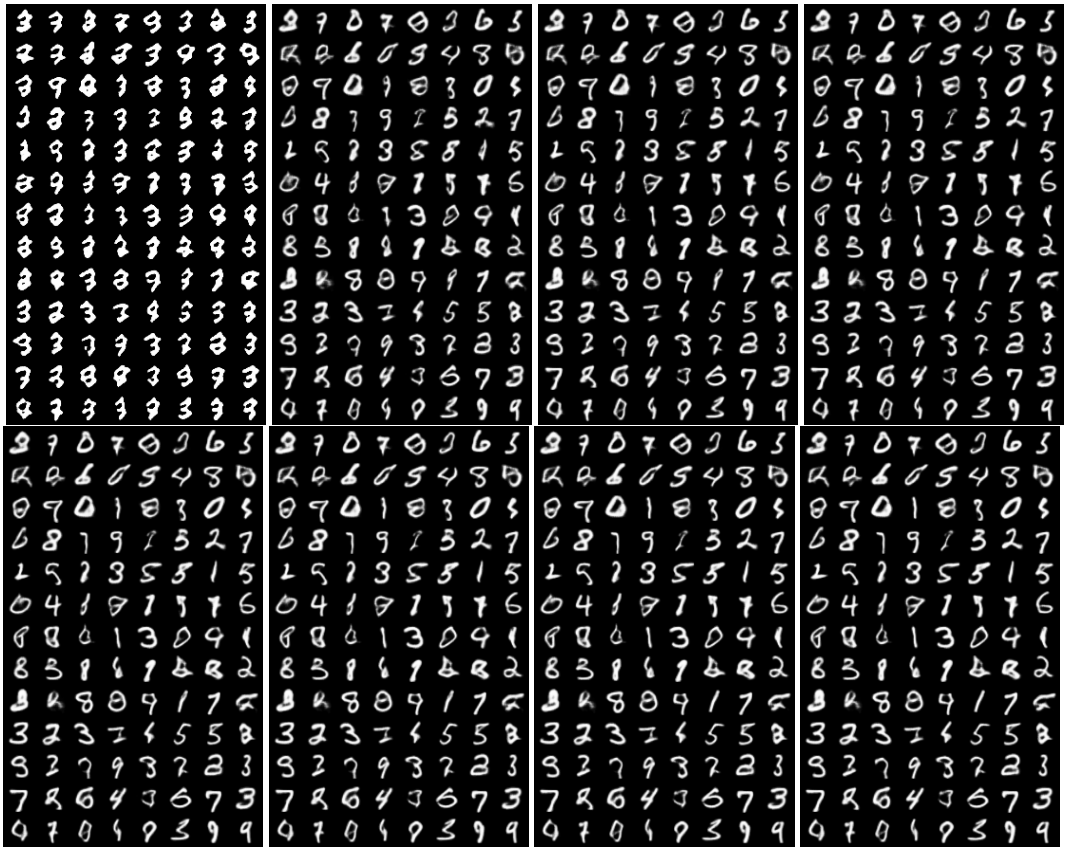


Figure S3. The evolution of SWF through 200 iterations on the MNIST dataset. Plots are for 1, 11, 21, 31, 41, 51, 101 and 201 iterations

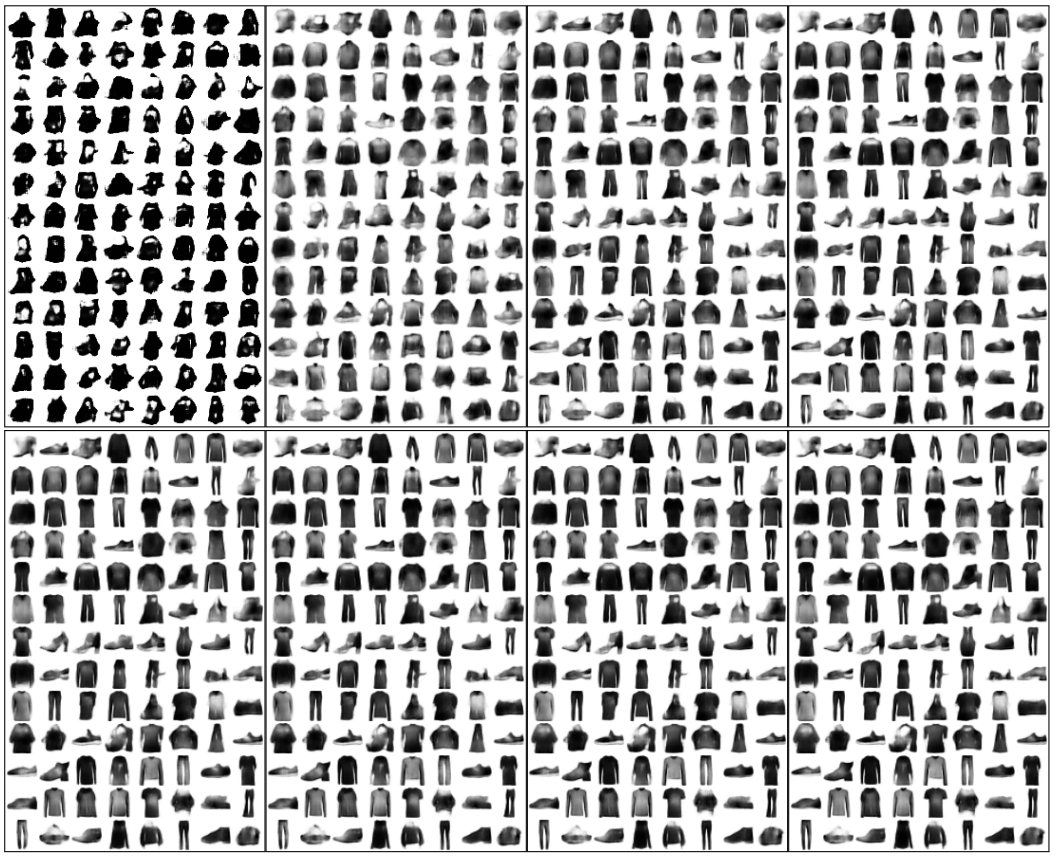
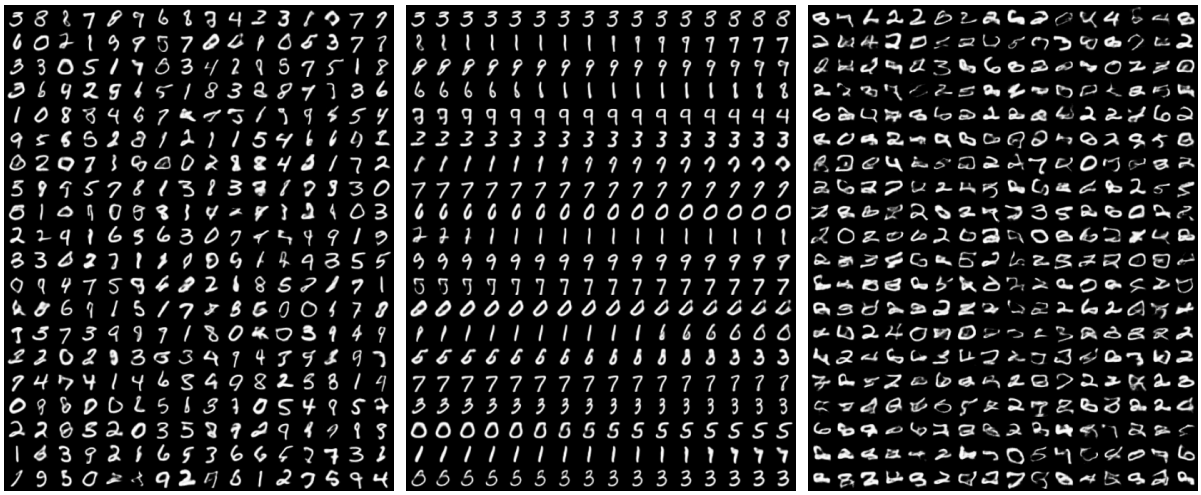


Figure S4. The evolution of SWF through 200 iterations on the FashionMNIST dataset. Plots are for 1, 11, 21, 31 (upper row) and 41, 51, 101, 201 (lower row) iterations



(a) particles undergoing SWF

(b) After SWF is done: applying learned map on linear combinations of train particles

(c) After SWF is done: applying learned map on random inputs.

Figure S5. SWF on MNIST: training samples, interpolation in learned mapping, extrapolation.

This noticed, we highlight that this generalization weakness of SWF for high dimensions is not really an issue, since it is always possible to i/ run SWF with more training samples if generalization is required ii/ re-run the algorithm for a set of new particles. Remember indeed that this does not require passing through the data again, since the distribution of the data projections needs to be done only once.

References

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

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Dalalyan, A. S. 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.

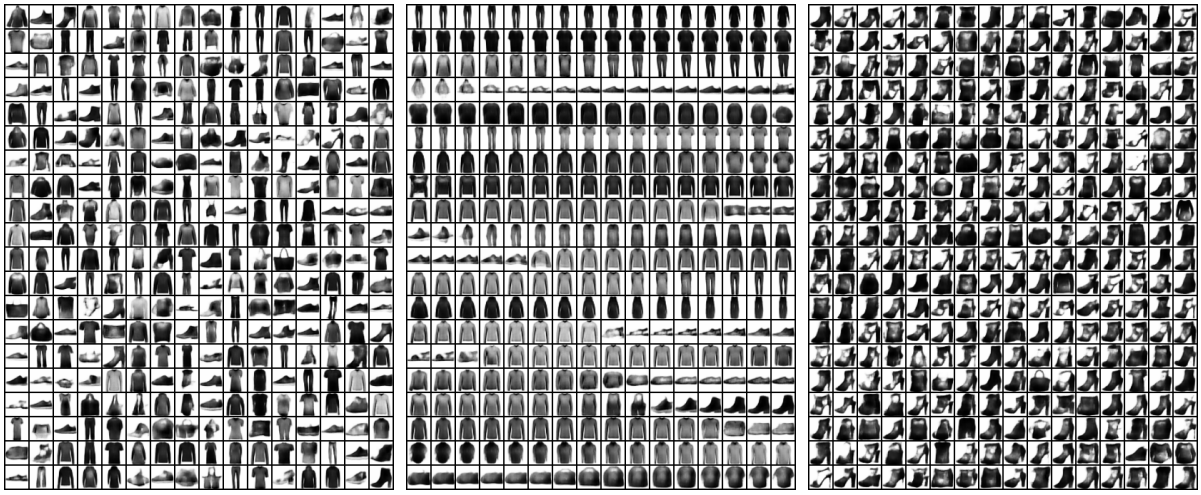
Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.



(a) particles undergoing SWF

(b) After SWF is done: applying learned map on linear combinations of train particles

(c) After SWF is done: applying learned map on random inputs.

Figure S6. SWF on FashionMNIST: training samples, interpolation in learned mapping, extrapolation.