Sobolev Descent

Youssef Mroueh[†] **Tom Sercu**[†] **Anant Raj**[⋆]
† IBM Research ⋆ MPI

Abstract

We study a simplification of GAN training: the problem of transporting particles from a source to a target distribution. Starting from the Sobolev GAN critic, part of the gradient regularized GAN family, we show a strong relation with Optimal Transport (OT). Specifically with the less popular *dynamic* formulation of OT that finds a path of distributions from source to target minimizing a "kinetic energy". We introduce Sobolev descent that constructs similar paths by following gradient flows of a critic function in a kernel space or parametrized by a neural network. In the kernel version, we show convergence to the target distribution in the MMD sense. We show in theory and experiments that regularization has an important role in favoring smooth transitions between distributions, avoiding large gradients from the critic. This analysis in a simplified particle setting provides insight in paths to equilibrium in GANs.

1 Introduction

We study the problem of transporting particles (cloud of high dimensional points) from a source to a target distribution, by incrementally following gradient flows of a critic function (Sobolev critic). We call this incremental process Sobolev Descent. This can be seen as a simplified version of GAN training dynamics: the generator is replaced by a set of N particles in $\mathbb{R}^d$. The particles define a time evolving distribution ν_{q_t}. Rather than min-max optimization in GANs, we only have maximization of the critic function f at each timestep t. We parametrize the critic either in an RKHS or with neural networks, leading us to Regularized Kernel and Neural Sobolev Descent respectively.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Optimal Transport (OT) [1, 2, 3] is increasingly gaining interest in the machine learning community. The static formulation of OT, seeks an optimal bijection T, defining a push forward operator from q to p: $T_{\#}\nu_q = \nu_p$ (i.e. Monge problem, relaxed by Kantorovic to seek a coupling π rather than a bijection T). While this static viewpoint is the most popular (e.g. WGAN [4] or recently [5, 6]), we will be focusing instead on the *dynamic* formulation of the Wasserstein-2 distance, for which Benamou and Brenier [7] showed that the OT problem has a fluid dynamic interpretation:

$$W_2^2(\nu_p, \nu_q) = \inf_{q_t, V_t} \int_0^1 \int \|V_t(x)\|^2 \, d\nu_{q_t}(x) dt$$
$$\text{s.t } \tfrac{\partial q_t(x)}{\partial t} = -div(q_t V_t(x)) \quad q_0 = q, q_1 = p. \qquad (1)$$

The optimal transport problem in this perspective corresponds to finding a *path of densities* q_t advecting from q to p with optimal *velocity fields* V_t that minimize the *kinetic energy*. Note that a major limitation is the need for an explicit analytic expression of p and q in order to solve for q_t, V_t in Eq. 1.

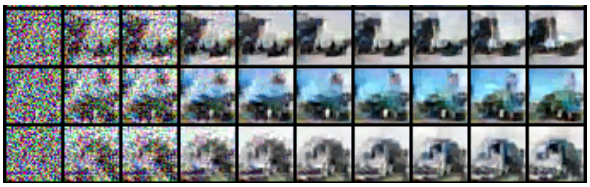

Figure 1: Neural Sobolev descent paths in the space of 32×32 images. The source distribution here is as in GAN uniform noise and the target is the truck class in CIFAR 10. The main difference with GAN is that images in Sobolev descent are the particles moving along the Sobolev critic, while in GAN the generator adapts in a min-max game with the critic.

In GANs, the generator is updated with stochastic gradient descent along directions of the discriminator (=critic) gradient $\nabla_x f(x)$, which immediately suggests a link with the velocity fields V_t in dynamic OT. In the recent GAN literature, variations have been studied where the gradient of the critic ($\mathbb{E}_{x\sim\mu}[\|\nabla f(x)\|]$) is constrained by adding a gradient penalty in the objective [8, 9]. We will show that for this specific class of critics, in the simplified particle descent setting, we construct

paths between source and target distributions that minimize a form of *kinetic energy*. Two advantages set it apart from dynamic OT: 1) we need only samples from p and q, and 2) the method is scalable (in sample size, input dimension and time complexity) because $f(x)$ is parametrized in an RKHS or with a neural network.

To define Sobolev descent we start from the recently introduced regularized Kernel Sobolev Discrepancy [10] as a way to quantify *the kinetic energy* that we wish to minimize (Section 2). We construct in Section 3 paths of distributions from source to target that minimize this notion of kinetic energy. We prove that under mild assumption kernel Sobolev descent converges in the MMD (Maximum Mean Discrepancy [11]) sense: $\text{MMD}(\nu_p, \nu_{q_t}) \to 0$ as $t \to \infty$. We highlight the prominent role of regularization in getting tunable smooth paths which relates to stable training in the GAN setting. We discuss the connections to dynamic OT [7] and Stein Descent of [12, 13] in Section 4. Finally in Section 5, we give algorithms for Kernel and Neural Sobolev Descent. We show the validity of our approach on synthetic data, image coloring and shape morphing and compare to classic OT algorithms. We then validate that Sobolev descent is a proxy for GANs on high dimensional data: we move particles $\in \mathbb{R}^{3\times32\times32}$ from noise to match CIFAR10 images (Figure 1).

2 Kernel Sobolev Discrepancy

In this Section we review the Kernel Sobolev Discrepancy recently introduced in [10]. The Kernel Sobolev Discrepancy will be fundamental to our particles descent as it defines the notion of kinetic energy to be minimized.

Sobolev Discrepancy. The Sobolev discrepancy was introduced recently in the context of Generative Adversarial Networks in Sobolev GAN [8]. We start by defining the Sobolev Discrepancy. Let $\mathcal{X}$ be a compact space in $\mathbb{R}^d$ with lipchitz boundary $\partial\mathcal{X}$.

Definition 1 (Sobolev Discrepancy [8, 10])**.** *Let ν_p, ν_q be two measures defined on $\mathcal{X}$. The Sobolev Discrepancy is defined as follows:*

$$\begin{aligned}\mathcal{S}(\nu_p, \nu_q) &= \sup_f \Big\{ \mathop{\mathbb{E}}_{x\sim\nu_p} f(x) - \mathop{\mathbb{E}}_{x\sim\nu_q} f(x) \Big\} \\ &\quad s.t\ f \in W_0^{1,2}(\mathcal{X}, \nu_q),\ \mathop{\mathbb{E}}_{x\sim\nu_q} \|\nabla_x f(x)\|^2 \le 1 \\ &= \inf_f \Big\{ \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 \, d\nu_q(x)} \Big\} \\ &\quad s.t\ p(x) - q(x) = -div(q(x)\nabla_x f(x)), f|_{\partial\mathcal{X}} = 0\end{aligned}$$

and $W_0^{1,2}(\mathcal{X}, \nu_q) = \{f$ vanishes at the boundary of $\mathcal{X}$ and $\mathop{\mathbb{E}}_{x\sim\nu_q} \|\nabla_x f(x)\|^2 < \infty\}$.

We refer to ν_p as the target distribution, and ν_q as the source distribution. The Sobolev discrepancy finds a witness function (or critic) that maximizes the mean discrepancy between the source and target distribution, while constraining the witness function gradients seminorm to be in a weighted Sobolev ball (under the source distribution ν_q). Note that the sup form (dual) is computationally friendly since it can be optimized using samples from p and q. The inf form (primal) sheds light on the physical meaning of this discrepancy: it is the **minimum kinetic energy** needed to advect the mass q to p following gradients of a critic. This interpretation will play a crucial role in Sobolev Descent.

Regularized Kernel Sobolev Discrepancy (RKSD). In order to define Sobolev descent we need to introduce a last ingredient: the Kernelized Sobolev Discrepancy. In other words a kernelized measure of *minimum kinetic energy for transporting q to p*. To simplify the presentation we give in the main paper results for finite dimensional RKHS, all results for infinite dimensional RKHS are given in Appendix C.

RKHS Properties and Assumptions. Let $\mathscr{H}$ be a *finite dimensional RKHS* with a finite feature map $\Phi : x \to \Phi(x) \in \mathbb{R}^m$, hence with Kernel k, $k(x,y) = \langle \Phi(x), \Phi(y) \rangle = \sum_{j=1}^m \Phi_j(x)\Phi_j(y)$, where $\langle , \rangle$ is the dot product in $\mathbb{R}^m$. Note that for a function $f \in \mathscr{H}$, $f(x) = \langle \boldsymbol{f}, \Phi(x) \rangle$, where $\boldsymbol{f} \in \mathbb{R}^m$ and $\|f\|_{\mathscr{H}} = \|\boldsymbol{f}\|$. Let $J\Phi(x) \in \mathbb{R}^{d\times m}$ be the jacobian of Φ, $[J\Phi]_{a,j}(x) = \frac{\partial}{\partial x_a}\Phi_j(x)$. We have the following expression of the gradient $\nabla_x f(x) = (J\Phi(x)\boldsymbol{f}) \in \mathbb{R}^d$. Mild assumptions on $\mathscr{H}$ are required (Φ bounded and differentiable (A1), has bounded derivatives (A2), and zero boundary condition on Φ (A3)) and can be found in [10].

Remark 1. *Assumption (A3) on zero boundary condition can be weakened to $q(x)\left\langle \nabla_x u^\lambda_{p,q}(x), n(x) \right\rangle = 0$ on $\partial\mathcal{X}$ ($n(x)$ is the normal on $\partial\mathcal{X}$). Assuming $\mathcal{X} = \mathbb{R}^d$ and that q and p vanish at ∞ we can use non vanishing feature maps Φ on $\partial\mathcal{X}$.*

The Kernel Sobolev Discrepancy [10] restricts the witness function of the Sobolev discrepancy to a finite dimensional RKHS $\mathscr{H}$, with feature map Φ. The Regularized Kernel Sobolev Discrepancy further regularizes the critic using Tikhonov regularization.

Definition 2 (RKSD)**.** *Let $\mathscr{H}$ be a finite dimensional RKHS satisfying assumptions A1, A2 and A3. Let $\lambda > 0$ be the regularization parameter. Let ν_p, ν_q be two measures defined on $\mathcal{X}$. The regularized Kernel Sobolev discrepancy restricted to the space $\mathscr{H}$ is defined*

as follows:

$$\mathcal{S}_{\mathscr{H},\lambda}(\nu_p,\nu_q) = \sup_{f\in\mathscr{H}} \Big\{ \mathop{\mathbb{E}}_{x\sim\nu_p} f(x) - \mathop{\mathbb{E}}_{x\sim\nu_q} f(x)\} \quad s.t \quad \mathop{\mathbb{E}}_{x\sim\nu_q} \|\nabla_x f(x)\|^2 + \lambda \|f\|^2_{\mathscr{H}} \leq 1 \tag{2}$$

We identify in the constraint in Equation (2) a regularized operator defined by

$$D(\nu_q) = \mathbb{E}_{x\sim\nu_q}([J\Phi(x)]^\top J\Phi(x)). \tag{3}$$

The constraint can be written as $\langle \boldsymbol{f}, (D(\nu_q)+\lambda I_m)\boldsymbol{f}\rangle \leq 1$. Following [10] we call $D(\nu_q)$ the Kernel Derivative Gramian Embedding (KDGE) of ν_q. KDGE is an operator embedding of the distribution.The KDGE can be seen as "covariance" of the jacobian. This operator embedding of ν_q is to be contrasted with the classic Kernel Mean Embedding (KME) of a distribution,

$$\boldsymbol{\mu}(\nu_q) = \mathbb{E}_{x\sim\nu_q}\Phi(x).$$

The KDGE can be thought as covariance of velocity fields (more on this intuition in Section 3.2).

The following proposition proved in [10] summarizes properties of the squared RKSD :

Proposition 1 (Closed Form Expression of RKSD)**.** *Let $\lambda > 0$. We have: $\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q) = \sup_{\boldsymbol{u}\in\mathbb{R}^m} 2\langle \boldsymbol{u}, \boldsymbol{\mu}(\nu_p)-\boldsymbol{\mu}(\nu_q)\rangle - \langle \boldsymbol{u}, (D(\nu_q)+\lambda I_m)\boldsymbol{u}\rangle$. This has the following closed form:*

$$\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q) = \left\| (D(\nu_q)+\lambda I_m)^{-\frac{1}{2}} (\boldsymbol{\mu}(\nu_p)-\boldsymbol{\mu}(\nu_q)) \right\|^2$$

and the optimal witness function $u^\lambda_{p,q}$ of $\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q)$ satisfies: $u^\lambda_{p,q}(x) = \langle \boldsymbol{u}^\lambda_{p,q}, \Phi(x)\rangle$ where

$$(D(\nu_q)+\lambda I_m)\boldsymbol{u}^\lambda_{p,q} = \boldsymbol{\mu}(\nu_p)-\boldsymbol{\mu}(\nu_q).$$

Note that $\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q) = \int_{\mathcal{X}} \|\nabla_x u^\lambda_{p,q}(x)\|^2 q(x)dx + \lambda \|\boldsymbol{u}^\lambda_{p,q}\|^2$ is the minimum regularized kinetic energy for advecting q to p using gradients of potentials in $\mathscr{H}$.

Note that RKSD is related to one of the most commonly used distances between distributions via embedding in RKHS, the maximum mean discrepancy [11]

$$\mathrm{MMD}(\nu_p,\nu_{q_t}) = \|\boldsymbol{\mu}(\mu_p)-\boldsymbol{\mu}(\nu_{q_t}\|,$$

with the main difference is that the KMEs in RKSD are whitened in the space defined by the KDGE defined in Eq. (3).

Remark 2. *a) From this proposition we see that $\nabla_x u^\lambda_{p,q}(x)$ can be seen as velocities of minimum regularized kinetic energy, advecting q to p. b) We give here the expression of the witness function $u^\lambda_{p,q}$ of $\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q)$ rather than $\mathcal{S}_{\mathscr{H},\lambda}(\nu_p,\nu_q)$ for convenience. The witness in (2) is $u^\lambda_{p,q}/\mathcal{S}_{\mathscr{H},\lambda}$.*

Empirical RKSD. An estimate of the Sobolev critic given finite samples from p and q $\{x_i, i=1\dots N, x_i\sim p\}$, and $\{y_i, i=1\dots M, y_i\sim q\}$ is straightforward: $\hat{u}^\lambda_{p,q}(x) = \langle \hat{\boldsymbol{u}}^\lambda_{p,q}, \Phi(x)\rangle_{\mathbb{R}^m}$, where $\hat{\boldsymbol{u}}^\lambda_{p,q} = (\hat{D}(\hat{\nu}_q)+\lambda I_m)^{-1}(\hat{\boldsymbol{\mu}}(\hat{\nu}_p)-\hat{\boldsymbol{\mu}}(\hat{\nu}_q))$. With the empirical KDGE is given by $\hat{D}(\hat{\nu}_q) = \frac{1}{M}\sum_{j=1}^M [J\Phi(y_j)]^\top J\Phi(y_j)$, and the emprical KMEs $\hat{\boldsymbol{\mu}}(\hat{\nu}_p) = \frac{1}{N}\sum_{i=1}^N \Phi(x_i)$ and $\hat{\boldsymbol{\mu}}(\hat{\nu}_q) = \frac{1}{M}\sum_{j=1}^M \Phi(y_j)$.

3 Sobolev Descent

Discrete Sobolev Descent. Now that we have a notion of Kernelized kinetic energy (the RKSD) and velocity fields consisting of the gradients of the Sobolev critic that achieve the minimum kinetic energy, we are ready to introduce the Sobolev Descent. Our main result will be to construct an infinitesimal transport map T^ε of the source distribution ν_q, and show that the resulting distribution $\nu_{q_{[T^\varepsilon]}}$ converges to the target distribution ν_p in the MMD sense. For $x\sim\nu_q$, moving along the gradient flow of the optimal regularized Sobolev critic $u^\lambda_{p,q}$ results in a decrease in MMD. We prove in Theorem 1 (all proofs are given in Appendix B) that, using the infinitesimal transport map:

$$T^\varepsilon(x) = x + \varepsilon\nabla_x u^\lambda_{p,q}(x),\ x\sim\nu_q,$$

the push forward $T^\varepsilon_\# \nu_q = \nu_{q_{[T^\varepsilon]}}$ ensures that this transport map decreases the MMD in the following sense:

$$\frac{d}{d\varepsilon}\mathrm{MMD}^2(\nu_p, T^\varepsilon_\#\nu_q)\Big|_{\varepsilon=0} \leq 0,$$

where the first variation $\frac{d}{d\varepsilon}\mathrm{MMD}^2(\nu_p,T^\varepsilon_\#\nu_q)|_{\varepsilon=0} = \lim_{\varepsilon\to 0}\frac{\mathrm{MMD}^2(\nu_p,T^\varepsilon_\#\nu_q)-\mathrm{MMD}^2(\nu_p,\nu_q)}{\varepsilon}$.

Theorem 1 (Gradient flows of the Regularized Sobolev Critic decrease the MMD distance)**.** *Let $\lambda>0$. Let $u^\lambda_{p,q}$ be the solution of the regularized Kernel Sobolev discrepancy between ν_p and ν_q i.e $\boldsymbol{u}^\lambda_{p,q} = (D(\nu_q)+\lambda I_m)^{-1}(\boldsymbol{\mu}(\nu_p)-\boldsymbol{\mu}(\nu_q))$. Consider the infinitesimal transport of ν_q via $T^\varepsilon(x) = x+\varepsilon\nabla_x u^\lambda_{p,q}(x)$. We have the following first variation of the MMD^2 under this particular perturbation:*

$$\frac{d}{d\varepsilon}MMD^2(\nu_p,T^\varepsilon_\#\nu_q)\Big|_{\varepsilon=0} = -2\left(MMD^2(\nu_p,\nu_q) - \lambda\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q)\right) \leq 0.$$

Remark 3. *1) The ≤ 0 of the RHS above is guaranteed since for any $\lambda>0$ we have $\mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p,\nu_q) \leq \left\|(D(\nu_q)+\lambda I)^{-1/2}\right\|^2_{op} \|\boldsymbol{\mu}(\nu_p)-\boldsymbol{\mu}(\nu_q)\|^2 \leq \frac{1}{\lambda}MMD^2(\nu_p,\nu_q)$ (where $\|.\|_{op}$ is the operator norm). 2) Assume $D(\nu_q)$ is non singular, Theorem 1 holds true for $\lambda=0$.*

From this Theorem we see that when we move the mass from q to p along the gradient flows of the regularized Sobolev critic, this results in a decrease in the MMD. Hence we are making progress towards matching p in the MMD sense. The amount of progress is proportional to $(\text{MMD}^2(\nu_p, \nu_q) - \lambda \mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p, \nu_q)) := \Delta_q$.

Theorem 1 suggests an iterative procedure that transports a source distribution ν_q to a target distribution ν_p: we start with applying transform $T_0^\varepsilon(x) = x + \varepsilon \nabla_x u^\lambda_{p,q_0}(x)$ on $q_0 = q$ which decreases the squared MMD distance by Δ_{q_0}. This results in a new distribution $q_1(x) = q_{0[T_0^\varepsilon]}(x)$. To further decrease the MMD distance we apply a new transform on q_1, $T_1^\varepsilon(x) = x + \varepsilon \nabla_x u^\lambda_{p,q_1}(x)$; this results in a decrease of the squared MMD distance by Δ_{q_1}. By iterating this process we construct a path of distributions $\{q_\ell\}_{\ell=0\dots L-1}$ between q_0 and p:

$$q_{\ell+1} = q_{\ell,[T_\ell^\varepsilon]} \text{ where } T_\ell^\varepsilon(x) = x + \varepsilon \nabla_x u^\lambda_{p,q_\ell}(x), x \sim \nu_{q_\ell}. \tag{4}$$

We call this iterative process Sobolev Descent, and this incremental decrease in the MMD distance is summarized in the following corollary:

Corollary 1 (Regularized Sobolev Descent Decreases the MMD)**.** *Consider the path of distributions q_ℓ between $q_0 = q$ and p constructed in equation (4) we have for $\ell \in \{0, \dots L-1\}$: $\frac{d}{d\varepsilon} MMD^2(\nu_p, \nu_{q_{\ell+1}})\Big|_{\varepsilon=0} = -2\left(MMD^2(\nu_p, \nu_{q_\ell}) - \lambda \mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p, \nu_{q_\ell})\right) \leq 0$.*

Continuous Sobolev Descent. We have refered to points advecting from q to p via Sobolev descent as particles. Let $t = \ell\varepsilon$ be the time variable, hence the time stepsize $dt = \varepsilon$. Note ν_{q_t} the measure of the moving particles X_t at time t. The continuous Sobolev descent can be defined at the limit $\varepsilon \to 0$ as the following non linear advection process on particles X_t (whose distribution is ν_{q_t}) advecting from q to p following the flow of the Sobolev critic :

$$dX_t = \nabla_x u_{p,q_t}(x) dt, X_0 \sim \nu_q,$$

where u_{p,q_t} is the Sobolev witness function between ν_p and ν_{q_t}. In the next section we will analyse the convergence of the continuous Sobolev descent to the target distribution ν_p.

3.1 Convergence of Continuous Sobolev Descent

In order to analyze the convergence of the continuous Sobolev descent, we will formulate the progress of MMD as a differential equation in time and show that the right hand side is always negative. There are two distinct cases.

Case 1: $\lambda = 0$, Unregularized Sobolev Discrepancy Flows. Assume that $D(\nu_{q_t})$ is non singular, for all time steps t. Corollary 1 suggests the following dynamic of the MMD for the continuous descent:

$$\frac{d}{dt}\text{MMD}^2(\nu_p, \nu_{q_t}) = -2\text{MMD}^2(\nu_p, \nu_{q_t}).$$

This suggests a fast exponential convergence of ν_{q_t} to ν_p in the MMD sense: $\text{MMD}^2(\nu_p, \nu_{q_t}) = e^{-2t}\text{MMD}^2(\nu_p, \nu_q)$, i.e $\text{MMD}^2(\nu_p, \nu_{q_t}) \to 0$, as $t \to \infty$. This fast convergence is not necessarily desirable as it may imply non-smooth paths with large discrete jumps from q_0 to p. For instance $q_t(x) = (1 - e^{-t})p(x) + e^{-t}q_0(x)$ exhibits this type of exponential convergence, but corresponds to intermediate distributions that are trivial interpolations between source and target distributions, and don't correspond to a meaningful smooth path from source to target, in the spirit of the Benamou-Brenier dynamic transport. See Figure 3 for an illustration.

Case 2: $\lambda > 0$ Regularized Sobolev Discrepancy Flows. In this case Corollary 1 suggests the following non linear dynamic of the MMD :

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\text{MMD}^2(\nu_p, \nu_{q_t}) &= -(\text{MMD}^2(\nu_p, \nu_{q_t}) - \lambda \mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p, \nu_{q_t})) \\ &\leq 0.\end{aligned}$$

Since $g(t) = \text{MMD}^2(\nu_p, \nu_{q_t})$ is decreasing and positive (bounded from below) it has a finite limit L as $t \to \infty$. When $g(t)$ reaches this limit at $t = t_0$ we have $\frac{dg(t)}{dt}|_{t=t_0} = 0$, and the graph of $g(t)$ remains constant, $g(t) = g(t_0) = L$ for $t \geq t_0$. Hence $\lim_{t\to\infty} \text{MMD}^2(\nu_p, \nu_{q_t}) = L = g(t_0)$.

We make here the following assumption on the target distribution ν_p that ensures that this limit L is zero.

Assumption (A): *For any measure ν_q, such that $\delta_{p,q} = \boldsymbol{\mu}(\nu_p) - \boldsymbol{\mu}(\nu_q) \neq 0$, $\delta_{p,q} \notin Null\,(D(\nu_q))$.* Assumption (A) means that we have: $D(\nu_q)(\boldsymbol{\mu}(\nu_p) - \boldsymbol{\mu}(\nu_q)) \neq 0$, for all q such that $\delta_{p,q} \neq 0$. This is a reasonable assumption and it is usually met in practice.

We show in Proposition 3 in Appendix B.1 that under assumption A, the regularized continuous Sobolev descent is convergent in the MMD sense: $\lim_{t\to\infty} \text{MMD}^2(\nu_p, \nu_{q_t}) = 0$.

Now if Assumption (A) does not hold, Sobolev Descent may stall at a $\nu_{q_{t_0}}$ where $\delta_{p,q_{t_0}} \in Null(D(\nu_{q_{t_0}}))$, and $\text{MMD}^2(\nu_p, \nu_{q_t}) \to \text{MMD}^2(\nu_p, \nu_{q_{t_0}}) \neq 0$ as $t \to \infty$.

Infinite dimensional RKHS, Characteristic kernel and Convergence in distribution of Sobolev Descent. For $\lambda > 0$, Theorem 1 holds true when Φ corresponds to an infinite dimensional feature map of

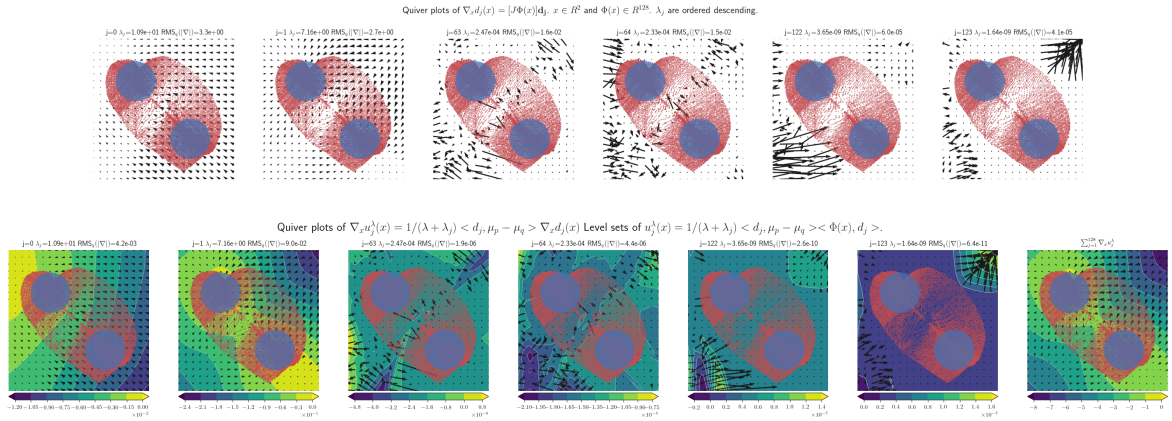

Figure 2: The principal transport directions for an intermediate state q_t (red cloud) in the shape morphing application with Neural Sobolev Descent (see Figure 6). The top row shows $\nabla_x d_j(x)$, bottom row shows $\nabla_x u_j^\lambda(x)$ for $\lambda = 0.3$. Note how small j (large eigenvalues) correspond to smooth vectorfields where the vectors have large norm (as measured in RMS over the points in point cloud $x \sim \nu_{q_t}$). The intermediate and large j values correspond to non-smooth vectorfields and non-smooth motions. For $\nabla_x u_j^\lambda(x)$, the principal transport directions $\nabla_x d_j(x)$ are multiplied with $\frac{1}{\lambda+\lambda_j}$ and the inner product with $\mu_p - \mu_q$ (two scalar multipliers). We see the non-smooth $\nabla_x u_j^\lambda(x)$ (small λ_j) have small RMS norm and contribute less, as they are effectively filtered out by the smoothing parameter λ. The bottom right subplot shows the total critic $u^\lambda(x) = \sum_{j=1}^m u_j^\lambda(x)$.

a characteristic kernel k, without any further assumptions (The case $\lambda = 0$ needs more care, and is tackled in Appendix C) . For $\lambda > 0$, under assumption (A) and for a characteristic kernel, the convergence of Sobolev descent in the MMD sense $\mathrm{MMD}(\nu_p, \nu_{q_t}) \to 0$ as $t \to \infty$, implies convergence in distribution : $\nu_{q_t} \xrightarrow{D} \nu_p$.

Damping effect of Regularization. While the MMD decreases at each time step, the regularization slows down the decrease of MMD by a factor proportional to the regularized Sobolev discrepancy $\lambda \mathcal{S}^2_{\mathscr{H},\lambda}(\nu_p, \nu_{q_t})$. Therefore regularization here is not only playing a computational role that stabilizes computation, it is also playing the role of a *damping*. This *damping* is desirable as it favors smoother paths between q_0 and p, i.e paths that deviates from the exponential regime in the un-regularized case. Hence we obtain tunable paths via regularization that favors smoother transitions from source to target (Fig 3).

3.2 Principal Transport Directions

In this section we shed light on how the flow of the Sobolev critic $\nabla_x u^\lambda_{p,q}(x)$ transports particles from q to p. To simplify notation we will omit the subscript p, q in this section but keep in mind that $u^\lambda(x)$ is to be determined for any intermediate state q_t. Let $(\lambda_j, \boldsymbol{d_j})$ be eigenvalues and eigenvectors of $D(\nu_q)$ (Eq 3) with $\lambda_j \geq 0$ descending. We can now think of $\nabla_x d_j(x)$ as **principal transport directions**, where $\nabla_x d_j(x) = [J\Phi(x)]\boldsymbol{d_j}$. This viewpoint becomes clear when we decompose the direction with which the particles move, i.e. the gradient of the critic $u^\lambda(x)$, over this basis $\nabla_x d_j(x)$. It is easy to see that: $u^\lambda(x) = \sum_{j=1}^m \frac{1}{\lambda_j+\lambda} \langle \boldsymbol{d_j}, \boldsymbol{\mu}(\nu_p) - \boldsymbol{\mu}(\nu_q) \rangle \langle \boldsymbol{d_j}, \Phi(x) \rangle$, and

$$\nabla_x u^\lambda(x) = \sum_{j=1}^m \nabla_x u_j^\lambda(x) = \sum_{j=1}^m \frac{1}{\lambda_j+\lambda} \langle \boldsymbol{d_j}, \boldsymbol{\mu}(\nu_p) - \boldsymbol{\mu}(\nu_q) \rangle \nabla_x d_j(x).$$

The Sobolev critic flow $\nabla_x u^\lambda(x)$ is thus decomposed on those principal transport directions, where each principal transport direction $\nabla_x d_j(x)$ is weighted by $\frac{1}{\lambda_j+\lambda} a_j$ where $a_j = \langle \boldsymbol{d_j}, \boldsymbol{\mu}(\nu_p) - \boldsymbol{\mu}(\nu_q) \rangle$. Let us first look at the meaning of a_j: if $a_j > 0$ this mean that this principal transport direction implies the correct motion advecting q to p (positively aligned with the difference of mean embeddings). On the other hand the term $\frac{1}{\lambda_j+\lambda}$, explains the role of regularization. Regularization is introducing a spectral filter on principal transport directions by weighing down directions with low eigenvalues. Those directions correspond to high frequency motions resulting in discrete jumps and discontinuous paths. Filtering them out with regularization parameter λ ensures smoother transitions in the probability path. See Figure 2 for an illustration. More in Appendix E.

3.3 Sobolev Descent as proxy for GANs

In this section we show how Sobolev descent can be seen as a proxy to GANs [14] that is more amenable to analysis. In Sobolev GAN [8], the critic between the current implicit distribution of the generator G_θ and the target distribution of real data $\mathbb{P}$ is updated. Then the generator is updated via gradient descent on the parameter space θ. This is similar to Sobolev descent, with the difference that GAN has a generator that is updated with the gradient flow of the critic, while Sobolev descent transports explicitly particles along that flow.

More formally Sobolev GAN [8] has the following updates: $f_t = \arg\max\{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim q_t} f(x) : f \in \mathscr{H}, \mathbb{E}_{x\sim q_t}||\nabla_x f(x)||^2 \leq 1\}$ where q_t is the distribution of the generator $G_{\theta_t}(z), z \sim p_z$. Using a continuous form of gradient descent on the generator parameter θ, we can write by the chain rule :

$$d\theta_t = \mathbb{E}_{\tilde{z}\sim p_z}\left[\frac{\partial G_\theta(\tilde{z})}{\partial\theta}\nabla_x f_t(G_\theta(\tilde{z}))\right]_{\theta=\theta_t} dt,$$

where $\frac{\partial G_\theta(\tilde{z})}{\partial\theta} \in \mathbb{R}^{|\theta|\times d}$ is the Jacobian matrix.

In order to match the particles intuition of Sobolev descent, we show here how to go from generator to particles. Fix z and set $X_t = G_{\theta_t}(z)$. X_t defines moving particles as θ_t is updated . Our goal is to see if the velocity of particles X_t produced by the generator in Sobolev GAN has similar behavior to the particles velocities in Sobolev descent. Using the chain rule we have : $dX_t = \frac{\partial G_\theta(z)}{\partial\theta}^\top\Big|_{\theta=\theta_t} d\theta_t$. Finally plugging the expression of $d\theta_t$ we have $dX_t =$

$$\mathbb{E}_{\tilde{z}\sim p_z}\Big(\frac{\partial G_\theta(z)}{\partial\theta}^\top \frac{\partial G_\theta(\tilde{z})}{\partial\theta}\Big|_{\theta=\theta_t}\nabla_x f_t(G_{\theta_t}(\tilde{z}))\Big)dt \quad (5)$$

If G_θ satisfies $(\frac{\partial G_\theta(z)}{\partial\theta}^\top \frac{\partial G_\theta(\tilde{z})}{\partial\theta}) = \delta(z-\tilde{z})I_d$ we recover the particles velocities of Sobolev descent: $dX_t = p(z)\nabla_x f_t(X_t)dt$, and our convergence analysis immediately applies to Sobolev GAN. Of course one needs to weaken the assumptions on G_θ to Lipschitzity of the Jacobian $\frac{\partial G_\theta(.)}{\partial\theta}$ in the latent space variable z and to carry further the analysis, we leave that for a future work. Our analysis of Sobolev descent suggests to consider gradient descent on the parameter space of the generator in GAN as a gradient flow on the probability space, corresponding to particles moving with a non linear Mckean Vlasov process [15] given in (5), and may allow under suitable conditions a theoretical understanding of GAN convergence complementing related works such as the ones of [16].

4 Relation to Previous Work

Dynamic OT of Benamou-Brenier [7] and Stein descent [13] are the closest to Sobolev Descent. The Benamou-Brenier formulation and Sobolev Descent minimize two related forms of kinetic energy in order to find paths connecting source and target distributions. Table 1 in Appendix F summarizes those main differences. In the Stein method [17, 18, 19, 20, 21], one of the measures ν_p is assumed to have a known density function p and we would like to measure the fidelity of samples from ν_q to the likelihood of p. The Stein discrepancy is obtained by applying a differential operator $T(p)$ to a vector valued function $\varphi : \mathcal{X} \to \mathbb{R}^d$, where $T(p)\varphi(x) = \langle\nabla_x \log(p(x)), \varphi(x)\rangle + \text{div}(\varphi(x))$. The Kernelized Stein Discrepancy is defined as follows: $\mathbb{S}(\nu_p, \nu_q) = \sup_\varphi\{\mathbb{E}_{x\sim\nu_q} T(p)\varphi(x) : \varphi_j \in \mathscr{H}, \sum_{j=1}^d \|\varphi_j\|^2_{\mathscr{H}} \leq 1\}$. Let $\varphi^*_{p,q}$ be the optimal solution. Variational Stein Descent of [13] uses $\varphi^*_{p,q}$ as a velocity field to transport particles distributed according to ν_q to approximate the target ν_p. This constructs paths reducing the KL divergence [12].

5 Algorithms and Experiments

Algorithms. We specify here the regularized Sobolev Descent for empirical measures $\hat{\nu}_p$ and $\hat{\nu}_q$,given finite samples from p and q: $\{x_i, i = 1\dots N, x_i \sim p\}$, and $\{y_i, i = 1\dots M, y_i \sim q\}$.

Empirical Regularized Kernel Sobolev Descent with Random Fourier Features. We consider the finite dimensional RKHS induced by random Fourier features [22]($\Phi(x) = \cos(Wx+b), W_{ij} \sim \mathcal{N}(0, 1/\sigma^2), b_i \sim \text{Unif}[0, 2\pi]$) . The empirical descent consists in using the estimate $\hat{u}^\lambda_{p,q}$ in Equation (4). For $\varepsilon > 0$, we have the following iteration, for $\ell \geq 1$ and all current positions of source particles $i = 1, \dots M$:

$$x_i^\ell = x_i^{\ell-1} + \varepsilon\nabla_x \hat{u}^\lambda_{p,q_{\ell-1}}(x_i^{\ell-1})$$

with $\hat{\nu}_{q_{\ell-1}}(dx) = \frac{1}{M}\sum_{i=1}^M \delta(x - x_i^{\ell-1})dx$, the empirical measure of particles $\{x_i^{\ell-1}, i = 1\dots M\}$, initialized with source particles $\{x_i^0 = y_i, i = 1\dots M\}$, and $\hat{u}^\lambda_{p,q_{\ell-1}}$ is the optimal critic of the empirical RKSD between empirical measure $\hat{\nu}_p$ and $\hat{\nu}_{q_{\ell-1}}$. The empirical regularized Kernel Sobolev Descent can be written as follows: for $l = \{1\dots L\}$:

$$\hat{\boldsymbol{u}}^\lambda_{p,q_{\ell-1}} = (\hat{D}(\hat{\nu}_{q_{\ell-1}}) + \lambda I_m)^{-1}\left(\hat{\boldsymbol{\mu}}(\hat{\nu}_p) - \hat{\boldsymbol{\mu}}(\hat{\nu}_{q_{\ell-1}})\right)$$

$$x_i^\ell = x_i^{\ell-1} + \varepsilon[J\Phi(x_i^{\ell-1})]\hat{\boldsymbol{u}}^\lambda_{p,q_{\ell-1}}, \forall i = 1, \dots M.$$

The Empirical Sobolev Descent is summarized in Algorithm 1, and the smoothness of the paths is controlled via the regularization parameter λ.

Neural Sobolev Descent. Inspired by the success of Sobolev GAN [8] that uses neural network approximations to estimate the Sobolev critic, we propose Neural Sobolev Descent. In Neural Sobolev Descent the critic function between ν_{q_t} and ν_p is estimated using a neural network $f_\xi(x) = \langle v, \Phi_\omega(x)\rangle$, where $\xi = (v, \omega)$ are the parameters of the neural network that we learn by gradient descent. We follow [8] in optimizing the parameters of the critic via an augmented Lagrangian. The particles descent is the same as in the Kernelized Sobolev Descent. Gradient descent on the parameters of the critic between updates of the particles resumes from the previous parameters (warm restart). Neural Sobolev Descent is summarized in Algorithm 2. Note that when compared to Sobolev GAN this descent replaces the generator with particles. It is worth

mentioning that regularization in the neural context is obtained via early stopping, i.e the number of updates n_c of the critic. Early stopping is known as a regularizer for gradient descent [23]. We will see that the smoothness of the paths is controlled via n_c. Note that GAN stabilization through early stopping (small critic updates) has been empirically observed [24, 25]. Our analysis suggests that this induces smoother paths for GANs.

Experiments. We confirm our theoretical findings on regularized Sobolev descent on a synthetic example highlighting the crucial role of regularization in smooth paths convergence. We then baseline Sobolev descent versus classical OT algorithms on the image color transfer problem. We show well-behaved trajectories of Sobolev descent in shape morphing thanks to smooth regularized paths.

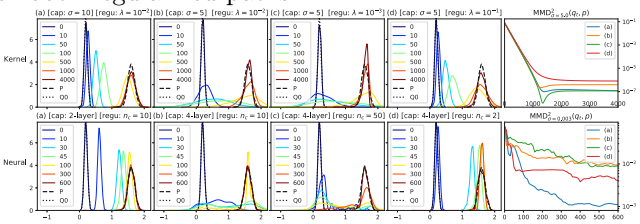

Figure 3: Moving 1000 samples of a 1D gaussian $\nu_{q_0} = \mathcal{N}(0.2, \sigma = 0.005)$ to $\nu_p = \mathcal{N}(1.6, \sigma = 0.1)$ with Kernel (top row) and Neural (bottom row) Sobolev Descent. Columns have similar properties between kernel and neural variants in terms of capacity of the model and regularization of the descent: (a) low capacity, (b) high capacity, (c) high capacity with decreased regularization and (d) high capacity with increased regularization.

Synthetic 1D Gaussians. Figure 3 shows Sobolev descent trajectories on a toy 1D problem, where both source and target are 1D Gaussians. Note that the Benamou-Brenier solution would be a *smooth* trajectory of normal distributions, where both the mean and standard deviation linearly interpolate between q_0 and p. Given 1000 samples from q_0 and p, we show in Figure 3 results of both kernel and neural Sobolev descent, where we plot kernel density estimators of densities at various time steps in the descent. We show the results of the descent for varying capacity of the function space (σ for the random features kernel, number of layers for Neural), and various regularization parameters (λ Tikhonov regularization for Kernel and n_c early stopping for Neural). Column (a) shows a regularized low capacity model achieving good approximation of the Benamou-Brenier optimal trajectory, where the data remains concentrated and smoothly moves to the target distribution. Column (b) shows a higher capacity model which blurs out the distribution before converging to p, where column (c) we even further decrease the regularization (smaller λ, bigger n_c) confirming the undesirable interpolation behavior which is predicted by the theory in the un-regularized case. Note that even in the Neural SD case this happens, corresponding to high frequency critic gradient behavior. In column (d) we increase the regularization on the high capacity model, achieving again a behavior that is closer to the optimal, without blurring or interpolation. This confirms the damping effect of regularization, filtering out the high frequency gradients. This can be also seen in the MMD plot in the last column. Figure 12 in Appendix H.2 gives similar results on morphing.

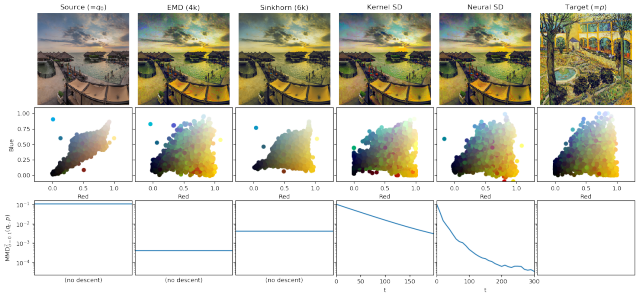

Figure 4: Color transfer. We compare Earth Mover Distance solved with linear programming on 4k samples, Sinkhorn on 6k samples with regularization $\varepsilon = 1e^{-2}$, Kernel SD with $\lambda = 1e^{-2}$ and $\sigma = 0.1$ at $t = 200$, and Neural SD at $t = 300$. The bottom row shows progress during the descent by computing the MMD(ν_{q_t}, ν_p) with bandwidth $\sigma = 0.1$ using 300 random Fourier features. Neural descent has a clear computational advantage over OT alternatives, which alleviates the need for subsampling and out of sample interpolation (which explains the high MMD values even for EMD).

Image Color Transfer. We consider the task of image color manipulation where we would like for an image A to match the color distribution of an image B. More formally, consider colored images Source and Target which we see as defining 3-dimensional probability distributions ν_{q_0} and ν_p, where every pixel is a sample: $\{x_1, x_2, \cdots x_N\} \sim \nu_{q_0}$ and $\{y_1, y_2, \cdots y_N\} \sim \nu_p$ and $N = 256 \times 256 = 66k$ the resolution. We move the samples using Kernel Sobolev Descent and Neural Sobolev Descent and analyze the distributions q_t. We provide in Figure 4 the results of our proposed algorithm on the task of image color transfer, comparing against results obtained with static Optimal Transport[1]. We show scatter plots after subsampling 5k points at random and display them on the (R,B) channels. In Appendix H Figures 10 and 11 we show the final MMD in function of rbf bandwidth σ and the evolution of the q_t distribution during the descent.

Shape Morphing with Sobolev Descent.

[1] We follow the recipe of [26] as implemented in the POT library [27] where we subsample for computational feasibility, then use interpolation for out-of-sample points.

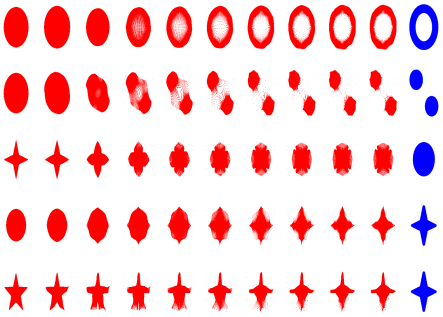

Figure 5: Morphing between several shapes using Kernelized Sobolev Descent. Intermediate steps are intermediate particles states of the Sobolev descent. Last column in the output of Kernelized Sobolev Descent.

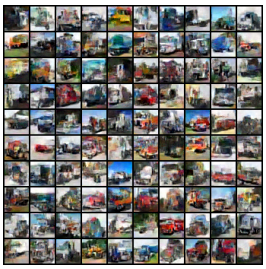

Figure 7: Particles (Images) of Neural Sobolev Descent at convergence, when the target distribution is the trucks class of CIFAR 10 and the Sobolev critic is a learned CNN.

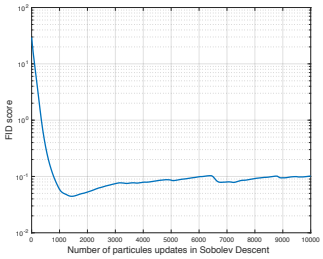

Figure 8: Frechet Inception Score (FID) of the particles produced by neural Sobolev descent as the descent progresses. FID is computed using the features from the second max pooling layer of the Inception v3 net (192-dim), by comparison against the truck class.

We use Sobolev descent for morphing between shapes. The source distribution is the distribution of points $x \in \mathbb{R}^2$ sampled uniformly from a shape A, that we need to move to become shape B. Such type of morphing has been considered in the Wasserstein Barycenter framework [28, 29]. Figure 5 shows the result of Kernelized Sobolev Descent (Algorithm 1) transforming between a source shape ν_q and a target shape ν_p, using random fourier features for $m = 100$ and $L = 600$, $\varepsilon = 0.01$ and $\lambda = 0.01$. We see that Kernelized Sobolev Descent morphs the shapes as the number of iterations

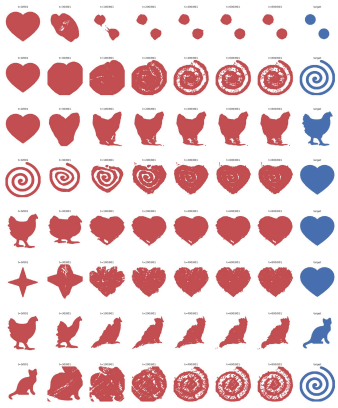

Figure 6: Morphing between several shapes using Neural Sobolev Descent. The descent is performed using a critic modeled by a simple 3-layer MLP.

approaches $L = 600$. Figure 6 shows Neural Sobolev Descent morphing between source shapes and target shapes. The first column is the source shape and last column is the target shape, in between columns are intermediate outputs of the Neural Sobolev Descent. Neural Sobolev Descent converges even on complex and unrelated shapes. Appendix H.3 provides the implementation and training details, and visualizes the critic $f_\xi(x)$ during the descent (Figure 13). Code is available on `https://bit.ly/2GtWXsY`. Videos of shapes morphing are available on `https://goo.gl/X4o8v6`.

High Dimensional Experiments: Transporting Noise to Images. We use neural Sobolev descent to transport uniform noise to the 5000 images in CIFAR10 labeled truck, similar to a typical GAN setup. The Sobolev critic architecture is a DCGAN discriminator architecture [30]. We see in Fig 7 that Sobolev descent converges and produces samples similar to the images from a trained GAN. The FID score [31] along the descent is given in Fig 8. This experiment confirms qualitatively and quantitatively our theoretical findings on Sobolev descent as a simplified proxy for GANs.

6 Conclusion

We introduced Sobolev descent on particles as a simplified proxy to GAN training. Sobolev descent constructs paths of distributions which minimize a kinetic energy, similar to dynamical Optimal Transport. We highlighted its convergence, its capacity in modeling high dimensional distributions and the crucial role of regularization in obtaining smooth transition paths by filtering out high frequency gradients. Our work sheds light on gradient based learning of GANs such as Sobolev GAN [8], that can be seen as a dynamic transport rather than the static as popularized by WGAN [4]. Our analysis explains GAN stabilization through early stopping (small updates of critic) [24, 25] as a regularization on the critic, inducing smoother paths to equilibrium.

References

[1] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2008.

[2] Filippo Santambrogio. Optimal transport for applied mathematicians. May 2015.

[3] Gabriel Peyré and Marco Cuturi. Computational optimal transport. Technical report, 2017.

[4] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein gan. *Arxiv*, 2017.

[5] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. 2018.

[6] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. Technical report, 2017.

[7] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 2000.

[8] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *ICLR*, 2018.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv:1704.00028*, 2017.

[10] Youssef Mroueh. Regularized finite dimensional kernel sobolev discrepancy. *Tech Rep. arXiv:1805.06441*, 2018.

[11] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

[12] Qiang Liu. Stein variational descent as a gradient flow. *NIPS*, 2017.

[13] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems 29*. 2016.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014.

[15] A. A. Vlasov. The vibrational properties of an electron gas. *Phys. Usp.*, 1968.

[16] Léon Bottou, Martín Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. In *Braverman Readings in Machine Learning*, volume 11100 of *Lecture Notes in Computer Science*, pages 229–268. Springer, 2017.

[17] Jackson Gorham and Lester W. Mackey. Measuring sample quality with stein's method. In *NIPS*, pages 226–234, 2015.

[18] Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016.

[19] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML 2016*, 2016.

[20] Jack Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester W. Mackey. Measuring sample quality with diffusions. *CoRR*, abs/1611.06972, 2016.

[21] Jackson Gorham and Lester W. Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

[22] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*. 2008.

[23] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 2007.

[24] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

[25] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease adivergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

[26] Sira Ferradans, Nicolas Papadakis, Julien Rabin, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 2013.

[27] Rémi Flamary and Nicolas Courty. Pot python optimal transport library. 2017.

[28] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 2015.

[29] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, 2014.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Arxiv*, 2015.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[32] Rémi Peyre. Comparison between w2 distance and h- norm, and localisation of wasserstein distance. 2016.

[33] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 2008.

[34] Tadahisa Funaki. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1984.