# Unbalanced Sobolev Descent

**Youssef Mroueh , Mattia Rigotti**
IBM Research AI
mroueh@us.ibm.com,mrg@zurich.ibm.com

## Abstract

We introduce Unbalanced Sobolev Descent (USD), a particle descent algorithm for transporting a high dimensional source distribution to a target distribution that does not necessarily have the same mass. We define the *Sobolev-Fisher* discrepancy between distributions and show that it relates to advection-reaction transport equations and the Wasserstein-Fisher-Rao metric between distributions. USD transports particles along gradient flows of the witness function of the Sobolev-Fisher discrepancy (advection step) and reweighs the mass of particles with respect to this witness function (reaction step). The reaction step can be thought of as a birth-death process of the particles with rate of growth proportional to the witness function. When the *Sobolev-Fisher witness* function is estimated in a Reproducing Kernel Hilbert Space (RKHS), under mild assumptions we show that USD converges asymptotically (in the limit of infinite particles) to the target distribution in the Maximum Mean Discrepancy (MMD) sense. We then give two methods to estimate the *Sobolev-Fisher witness* with neural networks, resulting in two Neural USD algorithms. The first one implements the reaction step with mirror descent on the weights, while the second implements it through a birth-death process of particles. We show on synthetic examples that USD transports distributions with or without conservation of mass faster than previous particle descent algorithms, and finally demonstrate its use for molecular biology analyses where our method is naturally suited to match developmental stages of populations of differentiating cells based on their single-cell RNA sequencing profile. Code is available at http://github.com/ibm/usd.

## 1 Introduction

Particle flows such as Stein Variational Gradient descent [1], Sobolev descent [2] and MMD flows [3], allow the transport of a source distribution to a target distribution, following paths that progressively decrease a discrepancy between distributions (Kernel Stein discrepancy and MMD, respectively). Particle flows can be seen through the lens of Optimal Transport as gradient flows in the Wasserstein geometry [4], and they've been recently used to analyze the dynamics of gradient descent in over-parametrized neural networks in [5] and of Generative Adversarial Networks (GANs) training [2].

Unbalanced Optimal Tansport [6, 7, 8, 9] is a new twist on the classical Optimal Transport theory [10], where the total mass between source and target distributions may not be conserved. The Wasserstein Fisher-Rao (WFR) distance introduced in [7] gives a dynamic formulation similar to the so-called Benamou-Brenier dynamic form of the Wasserstein-2 distance [11], where the dynamics of the transport is governed by an advection term with a velocity field $V_t$ and a reaction term with a rate of growth $r_t$, corresponding to the construction and destruction of mass with the same rate:

$$\mathrm{WFR}^2(p,q) = \inf_{q_t,V_t,r_t} \int_0^1 \int (\|V_t(x)\|^2 + \frac{\alpha}{2} r_t^2(x)) dq_t(x) dt$$
$$\text{s.t } \frac{\partial q_t(x)}{\partial t} = -\mathrm{div}(q_t(x)V_t(x)) + \alpha\, r_t(x) q_t(x), \qquad q_0 = q, q_1 = p. \tag{1}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

From a particle flow point of view, this advection-reaction in Unbalanced Optimal Transport corresponds to processes of birth and death, where particles are created or killed in the transport from source to target. Particle gradient descent using the WFR geometry have been used in the analysis of over-parameterized neural networks and implemented as Birth-Death processes in [12] and as conic descent in [13]. In the context of particles transportations, [14] showed that birth and death processes can accelerate the Langevin diffusion. On the application side, Unbalanced Optimal Transport is a powerful tool in biological modeling. For instance, the trajectories of a tumor growth have been modeled in the WFR framework by [15]. [16] and [17] used Unbalanced Optimal Transport to find differentiation trajectories of cells during development.

The dynamic formulation of WFR is challenging as it requires solving PDEs. One can use the unbalanced Sinkhorn divergence and apply an Euler scheme to find the trajectories between source and target as done in [18] but this does not give any convergence guarantees.

In this paper we take another approach similar to the one of Sobolev Descent [2]. We introduce the Kernel Sobolev-Fisher discrepancy that is related to WFR and has the advantage of having a closed form solution. We present a particle descent algorithm in the unbalanced case named *Unbalanced Sobolev Descent* (USD) that consists of two steps: an advection step that uses the gradient flows of a witness function of the Sobolev-Fisher discrepancy, and a reaction step that reweighs the particles according to the witness function. We show theoretically that USD is convergent in the Maximum Mean Discrepancy sense (MMD), that the reaction step accelerates the convergence, in the sense that it results in strictly steeper descent directions, and give a variant where the witness function is efficiently estimated as a neural network. We then empirically demonstrate the effectiveness and acceleration of USD in synthetic experiments, image color transfer tasks, and finally use it to model the developmental trajectories of populations of cells from single-cell RNA sequencing data [16].

## 2 Sobolev-Fisher Discrepancy

In this Section we define the Sobolev-Fisher Discrepancy (SF) and show how it relates to advection-reaction PDEs. While this formulation remains computationally challenging, we'll show in Section 3 how to approximate it in RKHS.

### 2.1 Advection-Reaction with no Conservation of Mass

**Definition 1** (Sobolev-Fisher Discrepancy)**.** *Let $p, q$ be two measures defined on $\mathcal{X} \subset \mathbb{R}^d$. For $\alpha > 0$, the Sobolev-Fisher Discrepancy is defined as follows:*

$$SF(p,q) = \sup_f \left\{ \mathbb{E}_{x\sim p} f(x) - \mathbb{E}_{x\sim q} f(x) : \quad \mathbb{E}_{x\sim q} \left\|\nabla_x f(x)\right\|^2 + \alpha \mathbb{E}_{x\sim q} f^2(x) \leq 1, \quad f|_{\partial\mathcal{X}} = 0 \right\}$$

Note that the objective of SF is an Integral Probability Metric (IPM) objective, and the function space imposes a constraint on the weighted Sobolev norm of the witness function $f$ on the support of the distribution $q$. We refer to $q$ as the source distribution and $p$ as the target distribution. The following theorem relates the solution of the Sobolev-Fisher Discrepancy to an advection-reaction PDE:

**Theorem 1** (Sobolev-Fisher Critic as Solution of an Advection-Reaction PDE)**.** *Let $u$ be the solution of the* advection-reaction *PDE:*

$$p(x) - q(x) = -div(q(x)\nabla_x u(x)) + \alpha u(x) q(x), \quad u|_{\partial\mathcal{X}} = 0.$$

*Then $SF^2(p,q) = \mathbb{E}_{x\sim q} \left\|\nabla_x u(x)\right\|^2 + \alpha \mathbb{E}_{x\sim q} u^2(x)$, with witness function $f^*_{p,q} = u/SF(p,q)$.*

From Theorem 1 we see that the witness function of $SF^2$ solves an advection-reaction where the mass is transported from $q$ to $p$, via an advection term following the gradient flow of $\nabla_x u$, and a reaction term amounting to construction/destruction of mass that we also refer to as a birth-death process with a rate given by $u$. Intuitively, if the witness function $u(x) > 0$ we need to create mass, and destruct mass if $u(x) < 0$. This is similar to the notion of particle birth and death defined in [12] and [14].

In Proposition 1 we give a convenient unconstrained equivalent form for $SF^2$:

**Proposition 1** (Unconstrained Form of $SF^2$)**.** *SF satisfies the expression: $SF^2(p,q) = \sup_u L(u)$, with $L(u) = 2(\mathbb{E}_{x\sim p} u(x) - \mathbb{E}_{x\sim q} u(x)) - \left( \mathbb{E}_{x\sim q} \left\|\nabla_x u(x)\right\|^2 + \alpha \mathbb{E}_{x\sim q} u^2(x) \right)$.*

Theorem 2 gives a physical interpretation for $\text{SF}^2$ as finding the witness function $u$ that has minimum sum of kinetic energy and rate of birth-death while transporting $q$ to $p$ via advection-reaction:

**Theorem 2** (Kinetic Energy & Birth-Death rates minimization)**.** *Consider the following minimization:*

$$P = \inf_{\substack{V:\mathcal{X}\to\mathbb{R}^d \\ r:\mathcal{X}\to\mathbb{R}}} \left\{ \frac{1}{2}\left( \int_{\mathcal{X}} (\|V(x)\|^2 + \alpha r^2(x)) q(x) dx \right) : p(x) - q(x) = -div(q(x)V(x)) + \alpha r(x) q(x) \right\}$$

*We then have that* $P = \frac{1}{2} SF^2(p, q)$*, and moreover:*

$$SF^2(p,q) = \inf_u \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x) dx + \alpha \int_{\mathcal{X}} u^2(x) q(x) dx,$$
$$\textit{subject to } p(x) - q(x) = -div(q(x)\nabla_x u(x)) + \alpha u(x) q(x).$$

**Remarks.** a) When $\alpha = 0$ we obtain the Sobolev Discrepancy, or $\|p - q\|_{\dot{H}^{-1}(q)}$, that linearizes the Wasserstein-2 distance. b) Note that this corresponds to a Beckman type of optimal transport [19], where we transport $q$ to $p$ ($q$ and $p$ do not have the same total mass) via an advection-reaction with mass not conserved. It is easy to see that $\int_{\mathcal{X}} (p(x) - q(x)) dx = \alpha \int_{\mathcal{X}} u(x) q(x) dx$.

### 2.2 Advection-Reaction with Conservation of Mass

Define the Sobolev-Fisher Discrepancy with conservation of mass: $\overline{\text{SF}}^2(p,q) = \sup_u L(u)$, where $L(u) = 2\left(\mathbb{E}_{x\sim p} u(x) - \mathbb{E}_{x\sim q} u(x)\right) - \left( \mathbb{E}_{x\sim q} \|\nabla_x u(x)\|^2 + \alpha \left( \mathbb{E}_{x\sim q} \left( u(x) - \mathbb{E}_{x\sim q}(u(x)) \right)^2 \right) \right)$. The only difference between the previous expression and $\text{SF}^2$ in Proposition 1 is that the variance of the witness function is kept under control, instead of the second order moment. Defining

$$\mathcal{E}(u) = \int_{\mathcal{X}} (\|\nabla_x u(x)\|^2 + \alpha (u(x) - \mathbb{E}_{x\sim q} u(x))^2) q(x) dx$$

one can similarly show that $\overline{\text{SF}}$ has the primal representation:

$$\overline{\text{SF}}^2(p,q) = \inf_u \left\{ \mathcal{E}(u) : p(x) - q(x) = -div(q(x)\nabla_x u(x)) + \alpha (u(x) - \mathbb{E}_{x\sim q} u(x)) q(x) \right\}.$$

Hence, we see that $\overline{\text{SF}}$ is the minimum sum of kinetic energy and variance of birth-death rate for transporting $q$ to $p$ following an advection-reaction PDE with conserved total mass. The conservation of mass comes from the fact that $\chi(x) = -div(q(x)\nabla_x u(x)) + \alpha(u(x) - \mathbb{E}_{x\sim q} u(x)) q(x)$ satisfies:

$$\int_{\mathcal{X}} (p(x) - q(x)) dx = \int_{\mathcal{X}} \chi(x) = 0.$$

## 3 Kernel Sobolev-Fisher Discrepancy

In this section we turn to the estimation of SF discrepancy by restricting the witness function to a Reproducing Kernel Hilbert Space (RKHS), resulting in a closed-form solution.

### 3.1 Estimation in Finite Dimensional RKHS

Consider the finite dimensional RKHS, corresponding to an $m$ dimensional feature map $\Phi$:

$$\mathcal{H} = \{ f \,|\, f(x) = \langle w, \Phi(x) \rangle \text{ where } \Phi : \mathcal{X} \to \mathbb{R}^m, w \in \mathbb{R}^m \}.$$

Define the kernel mean embeddings $\mu(p) = \mathbb{E}_{x\sim p} \Phi(x)$, $\mu(q) = \mathbb{E}_{x\sim q} \Phi(x)$, and $\delta_{p,q} = \mu(p) - \mu(q)$. Let $C(q) = \mathbb{E}_{x\sim q} \Phi(x) \otimes \Phi(x)$ be the covariance matrix and $D(q) = \mathbb{E}_{x\sim q} J\Phi(x)^\top J\Phi(x)$ be the Gramian of the Jacobian, where $[J\Phi(x)]_{a,j} = \frac{\partial \Phi_j(x)}{\partial x_a}$, $a = 1 \dots d, j = 1 \dots m$.

**Definition 2** (Regularized Kernel Sobolev-Fisher Discrepancy (KSFD))**.** *Let* $u \in \mathcal{H}$*, and let* $\lambda > 0$ *and* $\gamma \in \{0, 1\}$*, define:* $L_{\gamma,\lambda}(u) = 2(\mathbb{E}_{x\sim p} u(x) - \mathbb{E}_{x\sim q} u(x)) - \left( \mathbb{E}_{x\sim q}[\|\nabla_x u(x)\|^2 + \alpha(u(x) - \gamma \mathbb{E}_q u(x))^2] + \lambda \|u\|^2_{\mathcal{H}} \right)$*. The Regularized Kernel Sobolev-Fisher Discrepancy is defined as:*

$$SF^2_{\mathcal{H},\gamma,\lambda}(p,q) = \sup_{u\in\mathcal{H}} L_{\gamma,\lambda}(u).$$

*When* $\gamma = 0$ *this corresponds to the unbalanced case, i.e. birth-death with no conservation of total mass, while for* $\gamma = 1$ *we have birth-death with conservation of total mass.*

**Proposition 2** (Estimation in RKHS)**.** *The Kernel Sobolev-Fisher Discrepancy is given by:* $SF^2_{\mathcal{H},\gamma,\lambda}(p,q) = \langle u^{\lambda,\gamma}_{p,q}, \delta_{p,q} \rangle$*, where the critic* $u^{\lambda,\gamma}_{p,q} = (D(q) + \alpha C_\gamma(q) + \lambda I_m)^{-1}\delta_{p,q}$*, with* $C_\gamma(q) = C(q) - \gamma\mu(q)\mu(q)^\top$*. Let* $u^{\lambda,\gamma}_{p,q}(x) = \langle u^{\lambda,\gamma}_{p,q}, \Phi(x) \rangle$ *and* $\delta_{p,q}(x) = \langle \delta_{p,q}, \Phi(x) \rangle$*, then:* $\nabla_x u^{\lambda,\gamma}_{p,q}(x) = (D(q) + \alpha C_\gamma(q) + \lambda I_m)^{-1}\nabla_x \delta_{p,q}(x)$.

**Remarks.** a) For the unbalanced case $\gamma = 0$, we refer to $\text{SF}^2_{\mathcal{H},0,\lambda}$ as $\text{SF}^2_{\mathcal{H},\lambda}$. For the case of mass conservation $\gamma = 1$, refer to $\text{SF}^2_{\mathcal{H},1,\lambda}$ as $\overline{\text{SF}}^2_{\mathcal{H},\lambda}$. Note that $C_1(q) = \bar{C}(q) = C(q) - \mu(q)\mu(q)^\top$. b) A similar Kernelized discrepancy was introduced in [20], but not as an approximation of the Sobolev-Fisher discrepancy, nor in the context of unbalanced distributions and advection-reaction. c) For $\alpha = 0$ we obtain the kernelized Sobolev Discrepancy KSD of [2].

### 3.2 Kernel SF for Direct Measures

Consider direct measures $p = \sum_{i=1}^N a_i \delta_{x_i}$ and $q = \sum_{j=1}^n b_j \delta_{y_j}$ (with no conservation of mass we can have $\sum_i a_i \neq \sum_j b_j \neq 1$ ). An estimate of the Sobolev-Fisher critic is given by $\hat{u}^{\lambda,\gamma}_{p,q} = (\hat{D}(q) + \alpha\hat{C}_\gamma(q) + \lambda I_m)^{-1}(\hat{\mu}(p) - \hat{\mu}(q))$, where the empirical Kernel Mean Embeddings are $\hat{\mu}(p) = \sum_{i=1}^N a_i \Phi(x_i)$ and $\hat{\mu}(q) = \sum_{j=1}^n b_j \Phi(y_j)$. The empirical operator embeddings are given by $\hat{D}(q) = \sum_{j=1}^n b_j [J\Phi(y_j)]^\top J\Phi(y_j)$, and $\hat{C}_\gamma(q) = \sum_{j=1}^n b_j \Phi(y_j)\Phi(y_j)^\top - \gamma\hat{\mu}(q)\hat{\mu}(q)^\top$.

## 4 Unbalanced Continuous Kernel Sobolev Descent

Given the Kernel Sobolev-Fisher Discrepancy defined in the previous sections and its relation to advection-reaction transport, in this section we construct a Markov process that transports particles drawn from a source distribution to a target distribution. Note that we don't assume that the densities are normalized nor have same total mass.

### 4.1 Constructing the Continuous Markov Process

Given $\alpha, \lambda > 0, \gamma \in \{0, 1\}$ and $n$ weighted particles drawn from the source distribution : $q_0^n = q = \sum_{i=1}^n b_i \delta_{y_i}$, i.e $X_i^0 = y_i$ and $w_i^0 = b_i$. Recall that the target distribution is given by $p = \sum_{i=1}^N a_i \delta_{x_i}$. We define the following Markov Process that we name Unbalanced Kernel Sobolev Descent:

$$
\begin{aligned}
dX_t^i &= \nabla_x u^{\lambda,\gamma}_{p,q_t^n}(X_t^i)dt \quad \text{(advection step)} \\
dw_t^i &= \alpha(u^{\lambda,\gamma}_{p,q_t^n}(X_t^i) - \gamma\mathbb{E}_{q_t^{(n)}} u^{\lambda,\gamma}_{p,q_t^n}(x))w_t^i dt \quad \text{(reaction step)} \\
q_t^n &= \sum_{i=1}^n w_t^i \delta_{X_t^i},
\end{aligned} \tag{2}
$$

where $u^{\lambda,\gamma}_{p,q_t^n}$ is the critic of the Kernel Sobolev-Fisher discrepancy, whose expression and gradients are given in Proposition 2. We see that USD consists of two steps: the advection step that updates the particles positions following the gradient flow of the Sobolev-Fisher critic, and a reaction step that updates the weights of the particles with a growth rate proportional to that critic. This reaction step consists in mass construction or destruction, that depends on the confidence of the witness function. This can be seen as birth-death process on the particles, where the survival log probability of a particle is proportional to the critic evaluation on this particle.

### 4.2 Generator Expression and PDE in the limit of $n \to \infty$

Proposition 3 gives the evolution equation of a functional of the intermediate distributions $q_t^n$ produced in the descent, at the limit of infinite particles $n \to \infty$:

**Proposition 3.** *Let* $\Psi : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$*, be a functional on the probability space. Let* $q_t^n$ *be the distribution produced by USD at time* $t$*. Let* $q_t$ *be its limit as* $n \to \infty$*, we have:*

$$\partial_t \Psi[q_t] = (\mathcal{L}\Psi)[q_t],$$

*where* $\mathcal{L}\Psi(q) = \int \left\langle \nabla_x u^{\lambda,\gamma}_{p,q}(x), \nabla_x D_q \Psi(x) \right\rangle q(dx) + \alpha \int D_q \Psi(x)(u^{\lambda,\gamma}_{p,q}(x) - \gamma \mathbb{E}_q u^{\lambda,\gamma}_{p,q}) q(x)dx$. *Where the functional derivative* $D_\mu$ *is defined through first variation for a signed measure* $\chi$ $(\int \chi(x)dx = 0)$*:*

$$\int D_\mu \Psi(x)\chi(x)dx = \lim_{\varepsilon \to 0} \frac{\Psi(\mu + \varepsilon\chi) - \Psi(\mu)}{\varepsilon}.$$

In particular, the paths of USD in the limit of $n \to \infty$ satisfy the advection-reaction equation:

$$\partial_t q_t = -div(q_t \nabla_x u^{\lambda,\gamma}_{p,q_t}) + \alpha(u^{\lambda,\gamma}_{p,q_t} - \gamma \mathbb{E}_{q_t} u^{\lambda,\gamma}_{p,q}) q_t.$$

### 4.3 Unbalanced Sobolev Descent decreases the MMD.

The following Theorem shows that USD when the number of the particles goes to infinity decreases the MMD distance at each step, where: $\text{MMD}^2(p,q) = \|\mu(p) - \mu(q)\|^2$.

**Theorem 3** (Unbalanced Sobolev Descent decreases the MMD)**.** *Consider the paths* $q_t$ *produced by USD. In the limit of particles* $n \to \infty$ *we have*

$$\frac{1}{2}\frac{dMMD^2(p,q_t)}{dt} = -\left(MMD^2(p,q_t) - \lambda SF^2_{\mathcal{H},\gamma,\lambda}(p,q_t)\right) \leq 0. \tag{3}$$

In particular, in the regularized case $\lambda > 0$ with strict descent (i.e. $q_t \neq p$ implies $\text{MMD}^2(p,q_t) - \lambda \text{SF}^2_{\mathcal{H},\gamma,\lambda}(p,q_t) > 0$), USD converges in the MMD sense: $\lim_{t\to\infty} \text{MMD}^2(p,q_t) = 0$. Similarly to [2], strict descent is ensured if the kernel and the target distribution $p$ satisfy the condition: $\delta_{p,q} \notin \text{Null}(D(q) + \alpha C_\gamma(q)), \forall q \neq p$.

**USD Accelerates the Convergence.** We now prove a Lemma the can be used to show that Unbalanced Sobolev Descent has an acceleration advantage over Sobolev Descent [2].

**Lemma 1.** *In the regularized case* $\lambda > 0$ *with* $\alpha > 0$*, the Kernel Sobolev-Fisher Discrepancy* $SF_{\mathcal{H},\gamma,\lambda}$ *is strictly upper bounded by the Kernel Sobolev discrepancy* $\mathcal{S}_{\mathcal{H},\lambda}$*(i.e for* $\alpha = 0$*) [2]:*

$$SF^2_{\mathcal{H},\gamma,\lambda}(p,q) < \mathcal{S}^2_{\mathcal{H},\lambda}(p,q).$$

From Lemma 1 and Eq. (3), we see that USD ($\alpha > 0$), results in a larger decrease in MMD than SD [2] ($\alpha = 0$), resulting in a steeper descent. Hence, USD advantages over SD are twofold: 1) it allows unbalanced transport, 2) it accelerates convergence for the balanced and unbalanced transport.

**USD with Universal Infinite Dimensional Kernel.** While we presented USD with a finite dimensional kernel for ease of presentation, we show in Appendix D that all our results hold for an infinite dimensional kernel. For a universal or a characteristic kernel, convergence in MMD implies convergence in distribution (see [21, Theorem 12]). Hence, using a universal kernel, USD guarantees the weak convergence as $\text{MMD}(p,q_t) \to 0$.

### 4.4 Understanding the effect of the Reaction Step: Whitened Principal Transport Directions

In [2] it was shown that the gradient of the Sobolev Discrepancy can be written as a linear combination of principal transport directions of the Gramian of derivatives $D(q)$. Here we show that unbalanced descent leads to a similar interpretation in a whitened feature space thanks to the $\ell_2$ regularizer. Let $\tilde{\mathcal{H}}_q = \{f \,|f(x) = \left\langle v, \tilde{\Phi}_q(x) \right\rangle, \tilde{\Phi}_q(x) = (C_\gamma(q) + \frac{\lambda}{\alpha} I)^{-\frac{1}{2}} \Phi(x)\}$, $\tilde{\delta}_{p,q} = (C_\gamma(q) + \frac{\lambda}{\alpha} I)^{-\frac{1}{2}} \delta_{p,q}$ $\tilde{D}(q) = (C_\gamma(q) + \frac{\lambda}{\alpha} I)^{-\frac{1}{2}} D(q) (C_\gamma(q) + \frac{\lambda}{\alpha} I)^{-\frac{1}{2}}$, and let $v^{\lambda,\gamma}_{p,q} = (\tilde{D}(q) + \alpha I_m)^{-1} \tilde{\delta}_{p,q}$. It is easy to see that the critic of the SF can be written as: $u^{\lambda,\gamma}_{p,q}(x) = \left\langle u^{\lambda,\gamma}_{p,q}, \Phi(x) \right\rangle = \left\langle v^{\lambda,\gamma}_{p,q}, \tilde{\Phi}_q(x) \right\rangle$. Note that $\tilde{\Phi}_q$ is a whitened feature map and $\tilde{D}(q)$ is the Gramian of its derivatives. Let $\tilde{d}_j, \lambda_j$ be the eigenvectors and eigenvalues of $\tilde{D}(q)$. We have: $v^{\lambda,\gamma}_{p,q} = \sum_{j=1}^m \frac{1}{\lambda_j + \alpha} \tilde{d}_j \left\langle \tilde{d}_j, \tilde{\delta}_{p,q} \right\rangle$. Hence, we write the gradient of the Sobolev-Fisher critic as $\nabla_x u^{\lambda,\gamma}_{p,q}(x) = \sum_{j=1}^m \frac{1}{\lambda_j + \alpha} \left\langle \tilde{d}_j, \tilde{\delta}_{p,q} \right\rangle [J\tilde{\Phi}(x)] \tilde{d}_j = \sum_{j=1}^m \frac{1}{\lambda_j + \alpha} \left\langle \tilde{d}_j, \tilde{\delta}_{p,q} \right\rangle \nabla_x \tilde{d}_j(x)$, where $\tilde{d}_j(x) = \left\langle \tilde{d}_j, \tilde{\Phi}_q(x) \right\rangle$. This says that the mass is transported along a weighted combination of whitened principal transport directions $\nabla_x \tilde{d}_j(x)$. $\alpha$ introduces a damping of the transport as it acts as a spectral filter on the transport directions in the whitened space.

## 5 Discrete time Unbalanced Kernel and Neural Sobolev Descent

In order to get a practical algorithm in this Section we discretize the continuous USD given in Eq. (2). We also give an implementation parameterizing the critic as a Neural Network.

**Discrete Time Kernel USD.** Recall that the source distribution $q_0 = q = \sum_{j=1}^n b_j \delta_{y_j}$, note $w_j^0 = b_j$ and $x_j^0 = y_j, j = 1 \dots n$. The target distribution $p = \sum_{j=1}^N a_j \delta_{x_j}$, and assume for simplicity $\sum_{j=1}^N a_j = 1$. Let $\varepsilon > 0$, for $\ell = 1 \dots L$, for $j = 1 \dots n$, we discretize the advection step:

$$x_j^\ell = x_j^{\ell-1} + \varepsilon \nabla_x u^{\lambda,\gamma}_{p,q_{\ell-1}}(x_j^{\ell-1}).$$

Let $m_{\ell-1} = \sum_{j=1}^n w_j^{\ell-1} u^{\lambda,\gamma}_{p,q_{\ell-1}}(x_j^{\ell-1})$. For $\tau > 0$, similarly we discretize the reaction step as:

$$a_j^\ell = \log(w_j^{\ell-1}) + \tau(u^{\lambda,\gamma}_{p,q_{\ell-1}}(x_j^{\ell-1}) - \gamma m_{\ell-1}).$$

If $\gamma = 0$ (total mass not conserved) we define the reweighing as follows: $w_j^\ell = \exp(a_j^\ell)$ and if $\gamma = 1$ (mass conserved): $w_j^\ell = \exp(a_j^\ell) / \sum_{i=1}^n \exp(a_i^\ell)$, and finally : $q^\ell = \sum_{j=1}^n w_j^\ell \delta_{x_j^\ell}$.

**Neural Unbalanced Sobolev Descent.** Motivated by the use of neural network critics in Sobolev Descent [2], we propose a Neural variant of USD by parameterizing the critic of the Sobolev-Fisher Discrepancy as a Neural network $f_\xi$ trained via gradient descent with the Augmented Lagrangian Method (ALM) on the loss function of SF given in Definition 1. The re-weighting is defined as in the kernel case above. Neural USD with re-weighting is summarized in Algorithm 1 in Appendix B. Note that the re-weighting can also be implemented via a birth-death process as in [12]. In this variant, particles are duplicated or killed with a probability driven by the growth rate given by the critic. We give the details of the implementation as birth-death process in Algorithm 2 (Appendix B).

**Computational and Sample Complexities.** The computational complexity Neural USD is given by that of updating the witness function and particles by SGD with backprop, i.e. $O(N(T + B))$, where $N$ is the mini-batch size, $T$ is the training time, $B$ is the gradient computation time for particles update. $T$ corresponds to a forward and a backward pass through the critic and its gradient. The sample complexity for estimating the Sobolev Fisher critic scales like $1/\sqrt{N}$ similar to MMD [22].

## 6 Relation to Previous Work

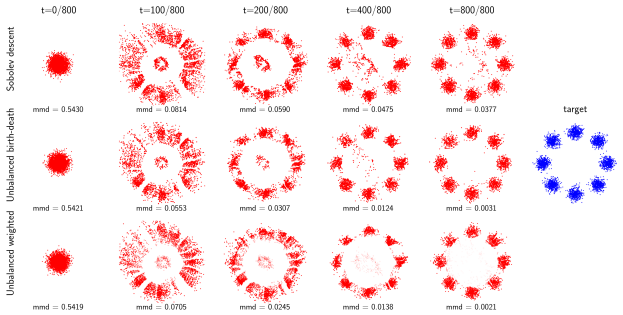
Table 1 in Appendix A summarizes the main differences between Sobolev descent [2], which only implements advection, and USD that also implements advection-reaction. Our work is related to the conic particle descent that appeared in [13] and [12]. The main difference of our approach is that it is not based on the flow of a fixed functional, but we rather learn dynamically the flow that corresponds to the witness function of the Sobolev-Fisher discrepancy. The accelerated Langevin Sampling of [14] also uses similar principles in the transport of distributions via Langevin diffusion and a reaction term implemented as a birth-death process. The main difference with our work is that in Langevin sampling the log likelihood of the target distribution is required explicitly, while in USD we only need access to samples from the target distribution. USD relates to unbalanced optimal transport [6, 7, 8, 9] and offers a computational flexibility when compared to Sinkhorn approaches [8, 9], since it scales linearly in the number of points while Sinkhorn is quadratic. Compared to WFR (Eq. (1)), USD finds greedily the connecting path, while WFR solves an optimal planning problem.

## 7 Applications

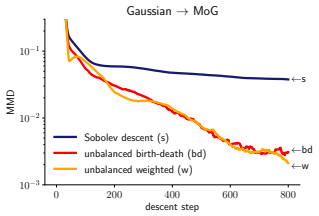
We experiment with USD on synthetic data, image coloring and prediction of developmental stages of scRNA-seq data. In all our experiments we report the MMD distance with a gaussian kernel, computed using the random Fourier features (RF) approximation [23] with 300 RF and kernel bandwith equal to $\sqrt{d}$ (the input dimension). We consider the conservation of mass case, i.e. $\gamma = 1$.

**Synthetic Examples.** We test Neural USD descent (Algorithms 1 and 2) on two synthetic examples. In the first example (Figure 1), the source samples are drawn from a 2D standard Gaussian, while target samples are drawn from a Mixture of Gaussians (MOG). Samples from this MOG have uniform

weights. In the second example (Figure 2), source samples are drawn from a 'cat'-shaped density whereas the target samples are drawn uniformly from a 'heart'. Samples from the targets have non-uniform weights following a horizontal gradient. In order to target such complex densities USD exploits advection and reaction by following the critic gradients and by creation and destruction of mass. We see in Figs 1 and 2 a faster mixing of USD in both, implementation with weights **(w)** and as birth-death **(bd)** processes compared to the Sobolev descent algorithm of [2].

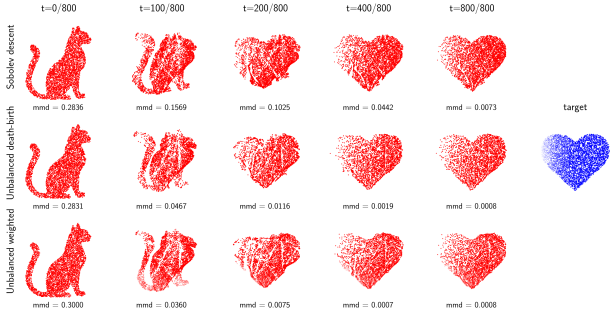



(a) Neural USD paths in transporting a Gaussian to a MOG. We compare Sobolev descent (SD, [2]) to both USD implementations: with birth-death process (bd: Algorithm 2) and weights (w: Algorithm 1). USD outperforms SD in capturing the modes of the MOG.

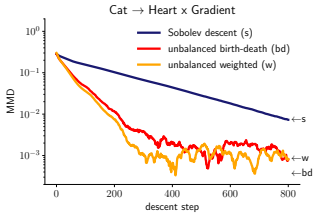



(b) MMD as a function of step along the descent from a Gaussian to a MOG. Both USD implementations convergence faster to the target distribution, reaching lower MMD than Sobolev Descent that relies on advection only.

Figure 1: Neural USD transport of a Gaussian to a MOG (target distribution is uniformly weighted).




(a) Neural USD transporting a 'cat' distributed cloud to a 'heart'. The main difference with the example above is that the points of the target distribution have non uniform weights describing a linear gradient as seen from the color code in the figure. Similarly to the MOG case, USD outperforms SD and better captures the non uniform density of the target.




(b) MMD as function of step along the descent from cat → heart × Grad. Similarly to the uniform target case USD accelerates the descent and outperforms SD.

Figure 2: Neural USD transport of a 'cat' to a non-uniform 'heart'. Samples from the target distribution have non-uniform weights given by $a_j$'s following a linearly decaying gradient.

**Image Color Transfer.** We test Neural USD on the image color transfer task. We choose target images that have sparse color distributions. This is a good test for unbalanced transport since intuitively having birth and death of particles accelerates the transport convergence in this case. We compare USD to standard optimal transport algorithms. We follow the recipe of [24] as implemented in the POT library [25], where images are subsampled for computational feasibility and then interpolated for out-of-sample points. We compare USD to Earth-Moving Distance (EMD), Sinkhorn [26] and Unbalanced Sinkhorn [8] baselines. We see in Figure 3 that USD achieves smaller MMD to the target color distribution. We give in Appendix H.2 in Fig 7 trajectories of the USD.

**Developmental Trajectories of Single Cells.** When the goal is not only to transport particles but also to find intermediate points along trajectories, USD becomes particularly interesting. This type of use case has recently received increased attention in developmental biology, thanks to single-cell RNA sequencing (scRNA-seq), a technique that records the expression profile of a whole population of cells at a given stage, but does so destructively. In order to trace the development of cells in-between such destructive measurements, [16] proposed to use unbalanced optimal transport [8].

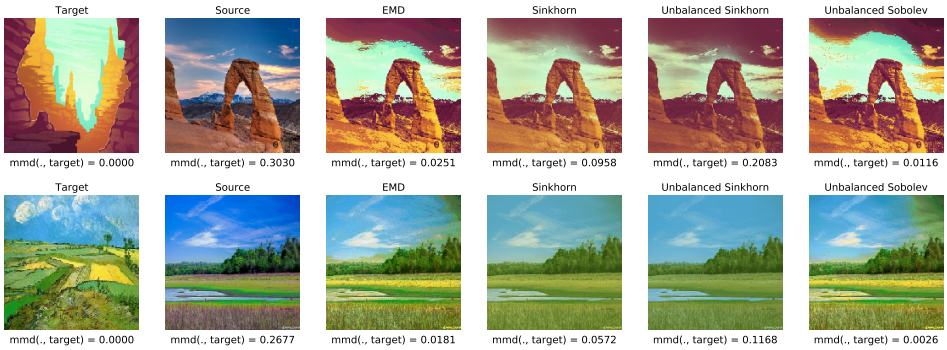



Figure 3: Color Transfer with USD using (bd) Algorithm 2. Comparison to OT baselines (EMD, Sinkhorn and Unbalanced Sinkhorn). USD achieves lower MMD, and faithfully captures the sparse distribution of the target.

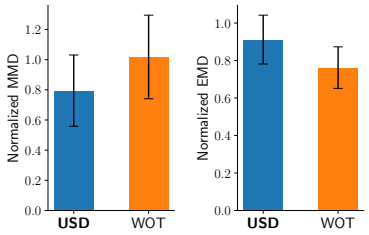



Figure 4: Mean and standard deviations plots of Normalized MMD and EMD for the intermediate stage prediction by USD and WOT (unbalanced OT) of [16] (means and standards deviation are computed over intervals). While USD outperforms WOT in MMD, the reverse holds in EMD. See text for an explanation.

Denoting those populations $q_{t_0}$ (source) and $q_{t_1}$ (target), then, in order to predict the population at an intermediate time $\frac{t_0+t_1}{2}$, [16] used a linear interpolation between matches between the source and target populations based on the coupling of unbalanced OT. This type of interpolation is a form of McCann interpolate [27]. As an alternative, we propose to use the mid-point of the USD descent as an interpolate, i.e. the timestamp in the descent $t_{1/2}$ such that $\mathrm{MMD}(q_{t_{1/2}}, q_{t_0}) = \mathrm{MMD}(q_{t_{1/2}}, q_{t_1})$. We test this procedure on the dataset released by [16]. For all time intervals $[t_0, t_1]$ in the dataset, we compute the intermediate stage $q_{t_{1/2}}$. We compare the quality of this interpolate with that obtained by the WOT algorithm of [16] in terms of MMD to the ground truth intermediate population $q^*_{t_{1/2}}$, normalized by MMD between initial and final population, i.e. $\mathrm{MMD}(q_{t_{1/2}}, q^*_{t_{1/2}})/\mathrm{MMD}(q_{t_0}, q_{t_1})$. Fig. 4 gives mean and standard deviation of the normalized MMD between intermediate stages predicted by USD and the ground truth. Note that mean and standard deviations are computed across 35 time intervals, individual MMDs can be found in Figure 8 in Appendix H. From Figure 4 we see that USD outperforms WOT in MMD, since USD is designed to decrease the MMD distance. On the other hand, for fairness of the evaluation we also report Normalized EMD (Earth-Mover Distance, normalized similarly) for which WOT outperforms USD. This is not surprising since WOT relies on unbalance OT, while USD instead provides guarantees in terms of MMD.

## 8 Conclusion

In this paper we introduced the KSFD discrepancy and showed how it relates to an advection-reaction transport. Using the critic of KSFD, we introduced Unbalanced Sobolev Descent (USD) that consists in an advection step that moves particles and a reaction step that re-weights their mass. The reaction step can be seen as birth-death process which, as we show theoretically, speeds up the descent compared to previous particle descent algorithms. We showed that the MMD convergence of Kernel USD and presented two neural implementations of USD, using weight updates, and birth and death of particle, respectively. We empirically demonstrated on synthetic examples and in image color transfer, that USD can be reliably used in transporting distributions, and indeed does so with accelerated convergence, supporting our theoretical analysis. As a further demonstration of our algorithm, we showed that USD can be used to predict developmental trajectories of single cells based on their RNA expression profile. This task is representative of a situation where distributions of different mass need to be compared and interpolated between, since the different scRNA-seq measurements are taken on cell populations of dissimilar size at different developmental stages. USD can naturally deal with this unbalanced setting. Finally we compared USD to unbalanced OT algorithms, showing its viability as a data-driven, more scalable dynamic transport method.

## Broader Impact Statement

Our work provides a practical particle descent algorithm that comes with a formal convergence proof and theoretically guaranteed acceleration over previous competing algorithms. Moreover, our algorithm can naturally handle situations where the objects of the descent are particles sampled from a source distribution descending towards a target distribution with different mass.

The type of applications that this enables range from theoretically principled modeling of biological growths processes (like tumor growth) and developmental processes (like the differentiation of cells in their gene expression space), to faster numerical simulation of advection-reaction systems.

Since our advance is mainly theoretical and algorithmic (besides the empirical demonstrations), its implications are necessarily tied to the utilization for which it is being deployed. Beside the applications that we mentioned, particle descent algorithms like ours have been proposed as a paradigm to characterize and study the dynamics of Generative Adversarial Network (GANs) training. As such, they could indirectly contribute to the risks associated with the nefarious uses of GANs such as deepfakes. On the other hand, by providing a tools to possibly analyze and better understand GANs, our theoretical results might serve as the basis for mitigating their abuse.

## Acknowledgments and Disclosure of Funding

Authors did not receive any third party funding and have no competing interests.

## References

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

[2] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Aistats,Proceedings of Machine Learning Research*, 2019.

[3] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *arXiv preprint arXiv:1906.04370*, 2019.

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

[5] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

[6] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[7] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[8] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[9] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

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

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

[12] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*, 2019.

[13] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

[14] Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

[15] Lénaïc Chizat and Simone Di Marino. A tumor growth model of hele-shaw type as a gradient flow, 2017.

[16] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[17] Karren D. Yang and Caroline Uhler. Scalable unbalanced optimal transport using generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[18] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

[19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *arXiv:1803.00567*, 2017.

[20] Michael Arbel, Dougal J. Sutherland, Mikolaj Binkowski, and Arthur Gretton. On gradient regularizers for mmd gans. *NeurIPS*, 2018.

[21] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions, 2016.

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

[23] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

[24] 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.

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

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

[27] Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 1997.

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

**Supplementary Material: Unbalanced Sobolev Descent**

### .1 Relation to Unbalanced Optimal Transport

We now relate our definition of the Sobolev-Fisher discrepancy to the following norm. For a signed measure $\chi$ define $\|\chi\|^2_{\dot{H}^{-1,2}(\nu)} =$

$$\sup_{f, \int_{\mathcal{X}}(\|\nabla_x f(x)\|^2 + \alpha f^2(x))d\nu \leq 1} \left| \int f d\chi \right| = \inf_{f, \chi(x) = -div(\nu(x)\nabla_x f(x)) + \alpha f(x)\nu(x)} \int_{\mathcal{X}} (\|\nabla_x f\|^2 + \alpha f^2) d\nu.$$

It can be shown that $\mathrm{SF}^2(p,q) = \|p-q\|^2_{\dot{H}^{-1,2}(q)}$.

The dynamic formulation of the Wasserstein Fisher-Rao metric given in Equation (1) can therefore be compactly written as:

$$\mathrm{WFR}^2(p,q) = \inf_{\nu_t} \int_0^1 \|d\nu_t\|^2_{\dot{H}^{-1,2}(\nu_t)} \, . \tag{4}$$

From this connection to WFR through $\|.\|^2_{\dot{H}^{-1,2}(q)}$, we see the link of the Sobolev-Fisher discrepancy to unbalanced optimal transport, since it linearizes the WFR for small perturbations.

## A Summary Table

| | $\alpha$ | $\gamma$ | Markov Process<br>Particles $j = 1 \dots n$ | PDE (As $n \to \infty$) | Guarantee<br>$\frac{1}{2}\frac{d\mathrm{MMD}^2(p,q_t)}{dt} =$ |
|---|---|---|---|---|---|
| **Sobolev Descent**<br>Flow of $\mathcal{S}_{\mathcal{H},\lambda}$<br>Target: $p = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$<br>Source : $q = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ | 0 | N/A | $dX_t^j = \nabla_x u^{\lambda}_{p,q_t}(X_t^j)dt$<br>$q_t = \frac{1}{n}\sum_{j=1}^n \delta_{X_t^j}$<br>Principal Transport Directions:<br>$dX_t = \sum_{\ell=1}^m \frac{1}{\lambda_\ell + \lambda} \langle d_\ell, \delta_{p,q_t} \rangle \nabla_x d_\ell(x) dt$<br>$(\lambda_j, d_j) = eig(D(q_t))$ | $\partial_t q_t = -div(q_t \nabla_x u^{\lambda}_{p,q_t})$<br>Advection | $-(\mathrm{MMD}^2(p,q_t) - \lambda \mathcal{S}^2_{\mathcal{H},\lambda}(p,q_t))$ |
| **Unbalanced Sobolev Descent:** Flow of $\mathrm{SF}_{\mathcal{H},\lambda}$<br>Target: $p = \sum_{i=1}^N a_i \delta_{x_i}$<br>Source : $q = \sum_{j=1}^n b_j \delta_{y_j}$<br>$(\sum_i a_i \neq \sum_j b_j)$ | $\alpha > 0$ | $\gamma = 0$ | $dX_t^j = \nabla_x u^{\lambda,\gamma}_{p,q_t}(X_t^j)dt$<br>$dw_t^j = \alpha(u^{\lambda,\gamma}_{p,q_t}(X_t^j))w_t^j dt$<br>$q_t = \sum_{i=1}^n w_t^i \delta_{X_t^i}$<br>Whitened Principal Transport Directions :<br>$dX_t^j = \sum_{\ell=1}^m \frac{1}{\lambda_\ell + \alpha} \left\langle \tilde{d}_\ell, \tilde{\delta}_{p,q} \right\rangle \nabla_x \tilde{d}_\ell(X_t^j) dt$ | $\partial_t q_t = -div(q_t \nabla_x u^{\lambda,\gamma}_{p,q_t}) + \alpha v^{\lambda,\gamma}_{p,q_t}(x) q_t$<br>Advection/Reaction<br>(Mass not conserved ) | $-(\mathrm{MMD}^2(p,q_t) - \lambda \mathrm{SF}^2_{\mathcal{H}}(p,q_t))$ |
| **Balanced Sobolev Descent:** Flow of $\overline{\mathrm{SF}}_{\mathcal{H},\lambda}$<br>Target: $p = \sum_{i=1}^N a_i \delta_{x_i}$<br>Source : $q = \sum_{j=1}^n b_j \delta_{y_j}$<br>$(\sum_i a_i = \sum_j b_j)$ | $\alpha > 0$ | $\gamma = 1$ | $dX_t^j = \nabla_x u^{\lambda,\gamma}_{p,q_t}(X_t^j)dt$<br>$dw_t^j = \alpha(u^{\lambda,\gamma}_{p,q_t}(X_t^j) - \mathbb{E}_{q_t} u^{\lambda,\gamma}_{p,q_t}) w_t^j dt$<br>$q_t = \sum_{i=1}^n w_t^i \delta_{X_t^i}$<br>Whitened Principal Transport Directions :<br>$dX_t^j = \sum_{\ell=1}^m \frac{1}{\lambda_\ell + \alpha} \left\langle \tilde{d}_\ell, \tilde{\delta}_{p,q} \right\rangle \nabla_x \tilde{d}_\ell(X_t^j) dt$ | $\partial_t q_t = -div(q_t \nabla_x u^{\lambda,\gamma}_{p,q_t}) + \alpha(u^{\lambda,\gamma}_{p,q_t}(x) - \mathbb{E}_{q_t} u^{\lambda,\gamma}_{p,q_t}) q_t$<br>Advection/Reaction<br>(Mass conserved ) | $-(\mathrm{MMD}^2(p,q_t) - \lambda \overline{\mathrm{SF}}^2_{\mathcal{H}}(p,q_t))$ |

Table 1: Summary table comparing Unbalanced Sobolev Descent to Sobolev Descent.

## B Algorithms

**Algorithm 3** CRITIC UPDATE($\xi$, target $\{(a_i, x_i)\}$, current source $\{(w_j^{\ell-1}, x_j^{\ell-1})\}$,$\gamma$)

**for** $j = 1$ **to** $n_c$ **do**
  $m_\xi \leftarrow \sum_{j=1}^n w_j^{\ell-1} f_\xi(x_j^{\ell-1})$
  $\hat{\mathcal{E}}(\xi) \leftarrow \sum_{i=1}^N a_i f_\xi(x_i) - m_\xi$
  $\hat{\Omega}(\xi) \leftarrow \sum_j w_j^{\ell-1} \left\| \nabla_x f_\xi(x_j^{\ell-1}) \right\|^2 + \alpha \left( \sum_j w_j^{\ell-1} f_\xi^2(x_j^{\ell-1}) - \gamma m_\xi^2 \right)$
  $\mathcal{L}_S(\xi, \lambda) = \hat{\mathcal{E}}(\xi) + \lambda(1 - \hat{\Omega}(\xi)) - \frac{\rho}{2}(\hat{\Omega}(\xi) - 1)^2$
  $(g_\xi, g_\lambda) \leftarrow (\nabla_\xi \mathcal{L}_S, \nabla_\lambda \mathcal{L}_S)(\xi, \lambda)$
  $\xi \leftarrow \xi + \eta$ ADAM $(\xi, g_\xi)$
  $\lambda \leftarrow \lambda - \rho g_\lambda$ {SGD rule on $\lambda$ with learning rate $\rho$}
**end for**
**Output:** $\xi$

**Algorithm 1** w-Neural Unbalanced Sobolev Descent (weighted version – ALM Algorithm)

**Inputs:** $\varepsilon, \tau$ Learning rate particles, $n_c$ number of critics updates, $L$ number of iterations, $\gamma \in \{0, 1\}$
$\{(a_i, x_i), i = 1 \dots N\}$, drawn from target distribution $\nu_p$
$\{(b_j, y_j), j = 1 \dots n\}$ drawn from source distribution $\nu_q$
Neural critic $f_\xi(x) = \langle v, \Phi_\omega(x) \rangle$, $\xi = (v, \omega)$ parameters of the neural network
**Initialize** $x_j^0 = y_j, w_j^0 = b_j$ for $j = 1 \dots n$
**for** $\ell = 1 \dots L$ **do**
  ***Critic Parameters Update***
  (between particles updates, gradient descent on the critic is initialized from previous episodes)
  $\xi \leftarrow$ CRITIC UPDATE($\xi$, target $\{x_i\}$, current source $\{(w_j^{\ell-1}, x_j^{\ell-1})\}$,$\gamma$ ) (Given in Alg. 3 in Appendix B)
  ***Particles and Weights Update***
  **for** $j = 1$ **to** $n$ **do**
    $x_j^\ell = x_j^{\ell-1} + \varepsilon \nabla_x f_\xi(x_j^{\ell-1})$ (current $f_\xi$ is the critic between $q_{\ell-1}$ and $p$, advection step)
    $a_j^\ell = \log(w_j^{\ell-1}) + \tau(f_\xi(x_j^{\ell-1}) - \gamma m_\xi)$ (reaction step)
    **if** $\gamma = 1$ (mass conservation) **then**
      $w^\ell = \mathrm{Softmax}(\mathrm{a}^\ell) \in \Delta_\mathrm{n}$
    **else if** $\gamma = 0$ (mass not conserved) **then**
      $w^\ell = \exp(a^\ell)$
    **end if**
  **end for**
**end for**
**Output:** $\{(x_j^L, w_j^L), j = 1 \dots n\}$

**Algorithm 2** bd-Neural Unbalanced Sobolev Descent (Birth-Death – ALM Algorithm)

**Inputs:** Same inputs of Algorithm 1
**Initialize** $x_j^0 = y_j, w_j^0 = \frac{1}{n}$ for $j = 1 \dots n$
**for** $\ell = 1 \dots L$ **do**
  ***Critic Parameters Update***
  (between particles updates gradient descent on the critic is initialized from previous episodes)
  $\xi \leftarrow$ CRITIC UPDATE($\xi$, target $\{(a_i, x_i)\}$, current source $\{(\frac{1}{n}, x_j^{\ell-1})\}$,$\gamma$ ) (Given in Alg. 3 in App. B)
  ***Particles and Weights Update (birth-death)***
  **for** $j = 1$ **to** $n$ **do**
    $x_j^\ell = x_j^{\ell-1} + \varepsilon \nabla_x f_\xi(x_j^{\ell-1})$ (current $f_\xi$ is the critic between $q_{\ell-1}$ and $p$ )
    $m_\xi \leftarrow \frac{1}{n} \sum_{i=1}^{j} f_\xi(x_i^\ell) + \frac{1}{n} \sum_{i=j+1}^{n} f_\xi(x_i^{\ell-1})$
    **if** $\beta_j = f_\xi(x_j^\ell) - \gamma m_\xi > 0$ **then**
      Duplicate $x_j^\ell$ with probability $1 - \exp(-\alpha\tau\beta_j)$
    **else if** $\beta_j = f_\xi(x_j^\ell) - \gamma m_\xi < 0$ **then**
      kill $x_j^\ell$ with probability $1 - \exp(-\alpha\tau|\beta_j|)$
    **end if**
  **end for**{Make population size $n$ again}
  $n_\ell$ number of particles at the end of the loop
  **if** $n_\ell > n$ **then**
    Kill $n_\ell - n$ randomly selected particles
  **else if** $n_\ell < n$ **then**
    Duplicate $n - n_\ell$ randomly selected partciles
  **end if**
**end for**
**Output:** $\{(x_j^L), j = 1 \dots n\}$

## C Proofs

*Proof of Theorem 1.* Define the following dot product between $u, v$ in the the Sobolev Space:

$$\langle u, v\rangle_{W_0^2} = \int_{\mathcal{X}} \langle \nabla_x u(x), \nabla_x v(x)\rangle \, q(x) + \alpha \int_{\mathcal{X}} u(x)v(x)q(x)dx,$$

and the norm :

$$\|u\|^2_{W_0^2} = \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 \, q(x)dx + \alpha \int_{\mathcal{X}} u^2(x)q(x)dx,$$

Let $f$ be any function such that $f|_{\partial\mathcal{X}=0}$, and $\|f\|_{W_0^2} \leq 1$:

$$\begin{aligned}
\mathcal{E}(f) &= \int_{\mathcal{X}} f(x)(p(x) - q(x))dx \\
&= -\int_{\mathcal{X}} f(x) div(q(x)\nabla_x u(x))dx + \alpha \int_{\mathcal{X}} u(x)f(x)q(x) \\
&= \int_{\mathcal{X}} \langle \nabla_x f(x), \nabla_x u(x)\rangle \, q(x) + \alpha \int_{\mathcal{X}} u(x)f(x)q(x)dx \\
&= \langle u, f\rangle_{W_0^2} \text{ (By definition)} \\
&\leq \|u\|_{W_0^2} \|f\|_{W_0^2} \text{ (By Cauchy Schwarz) ,} \\
&\leq \|u\|_{W_0^2} \text{ ($f$ feasible, } \|f\|_{W_0^2} \leq 1)
\end{aligned}$$

Let $f^*_{p,q} = u/\|u\|_{W_0^2}$, we have $\left\|f^*_{p,q}\right\|_{W_0^2} = 1$ and hence feasible, and it is easy to see that :

$$\mathcal{E}(f^*_{p,q}) = \|u\|_{W_0^2} ,$$

and hence we have that for all $f$ feasible we have:

$$\mathcal{E}(f) \leq \mathcal{E}(f^*_{p,q}),$$

and hence $f^*_{p,q}$ achieves the sup. □

*Proof of Proposition 1.* This can be easily proved using that $u^*$ solution of the PDE with source term is solution of that sup problem. $L(u^*) = \text{SF}^2(p, q)$ is clear from definition of $u^*$ we are left showing $L(u) \leq L(u*)$ for all $u$, this can be shown by proving that :

$$L(u) - L(u^*) = -\|u - u^*\|^2_{W_0^2} \leq 0$$

and hence $L(u) \leq L(u^*)$ , hence $u^*$ achieves the sup. □

*Proof of Theorem 2.* Writing the Lagrangian $u$ we have:

$$\inf_{V,r} \sup_{u} \mathcal{L}(V, r, u) = \sup_{u} \inf_{V,r} \mathcal{L}(V, r, u),$$

where By convexity of the cost we exchange sup and inf for $\mathcal{L}(V, r, u) = \frac{1}{2}\int_{\mathcal{X}} \|V(x)\|^2 \, q(x)dx + \alpha\frac{1}{2}\int_{\mathcal{X}} r^2(x)q(x)dx + \int_{\mathcal{X}} u(x)(p(x) - q(x)) - \int_{\mathcal{X}} \langle \nabla_x u(x), V(x)\rangle \, q(x) - \alpha \int_{\mathcal{X}} r(x)u(x)q(x)$. Note that $\inf_V \int_{\mathcal{X}} \|V(x)\|^2 \, q(x)dx - \int_{\mathcal{X}} \langle \nabla_x u(x), V(x)\rangle \, q(x) = -\sup_V \int_{\mathcal{X}} \langle \nabla_x u(x), V(x)\rangle \, q(x) - \frac{1}{2}\int_{\mathcal{X}} \|V(x)\|^2 \, q(x)dx = -\frac{1}{2}\int_{\mathcal{X}} \|\nabla_x u(x)\|^2 \, q(x)dx$(Fenchel Convex). Similarly we have: $\inf_r \frac{1}{2}\int_{\mathcal{X}} r^2(x)q(x)dx - \int_{\mathcal{X}} r(x)u(x)q(x) = -\sup_r \int_{\mathcal{X}} r(x)u(x)q(x) - \frac{1}{2}\int_{\mathcal{X}} r^2(x)q(x)dx = -\frac{1}{2}\int_{\mathcal{X}} u^2(x)q(x)dx$. Hence the dual problem is :

$$P = \sup_u \int_{\mathcal{X}} u(x)(p(x) - q(x))dx - \frac{1}{2}\left(\int_{\mathcal{X}} \|\nabla_x u(x)\|^2 \, q(x)dx + \alpha \int_{\mathcal{X}} u^2(x)q(x)dx\right)$$

By Proposition 1 ,we have :

$$P = \frac{1}{2}\text{SF}^2(p, q)$$

Hence $\text{SF}^2(p,q)$ has the equivalent form :

$$\text{SF}^2(p,q) = \inf_{V,r} \int_{\mathcal{X}} \|V(x)\|^2 q(x)dx + \alpha \int_{\mathcal{X}} r^2(x)q(x)dx$$
$$\text{Subject to: } p(x) - q(x) = -div(q(x)V(x)) + \alpha r(x)q(x)$$

Since $V^* = \nabla_x u$ and $r^* = u$ we have finally:

$$\text{SF}^2(p,q) = \inf_{u} \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x)dx + \alpha \int_{\mathcal{X}} u^2(x)q(x)dx$$
$$\text{Subject to: } p(x) - q(x) = -div(q(x)\nabla_x u(x)) + \alpha u(x)q(x).$$

□

*Proof of Proposition 2.*

$$\begin{aligned}
L_{\gamma,\lambda}(u) &= 2(\mathbb{E}_{x\sim p}u(x) - \mathbb{E}_{x\sim q}u(x)) - \left(\mathbb{E}_{x\sim q}[\|\nabla_x u(x)\|^2 + \alpha(u(x) - \gamma\mathbb{E}_q u(x))^2] + \lambda\|u\|^2_{\mathcal{H}}\right) \\
&= 2\langle u, \mu(p) - \mu(q)\rangle_{\mathcal{H}} - \left(\langle u, D(q)u\rangle_{\mathcal{H}} + \alpha(\mathbb{E}_q u^2(x) - \gamma(\mathbb{E}_{x\sim q}u(x))^2) + \lambda\|u\|^2_{\mathcal{H}}\right) \\
&= 2\langle u, \mu(p) - \mu(q)\rangle_{\mathcal{H}} - \left(\langle u, D(q)u\rangle_{\mathcal{H}} + \alpha(\langle u, C(q)u\rangle_{\mathcal{H}} - \gamma(\langle u, \mu(q)\rangle_{\mathcal{H}})^2 + \lambda\|u\|^2_{\mathcal{H}}\right) \\
&= 2\langle u, \mu(p) - \mu(q)\rangle_{\mathcal{H}} - \langle u, (D(q) + \alpha(C(q) - \gamma\mu(q)\otimes\mu(q)) + \lambda I)\, u\rangle_{\mathcal{H}}
\end{aligned}$$

Setting first order optimality for the sup we obtain:

$$(D(q) + \alpha(C(q) - \gamma\mu(q)\otimes\mu(q)) + \lambda I)\, u^{\lambda,\gamma}_{p,q} = \mu(p) - \mu(q) = \delta_{p,q}.$$

□

*Proof of Proposition 3.* For simplicity we give here the proof for $\gamma = 1$. $\gamma = 0$ has a similar proof. The proof follows ideas from [12]. Let $\Psi$ be a measure valued functional $\Psi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. For a measure $\mu$, $\Psi(\mu) \in \mathbb{R}$. The functional derivative $D_\mu$ is defined through first variation for a signed measure $\chi$ ($\int \chi(x)dx = 0$):

$$\int D_\mu \Psi(x)\chi(x)dx = \lim_{\varepsilon\to 0} \frac{\Psi(\mu + \varepsilon\chi) - \Psi(\mu)}{\varepsilon}$$

A generator function is defined as follows for a measure valued markov process $\mu_t^{(n)}$ (defined with $n$ particles) is defined as follows:

$$(\mathcal{L}_n\Psi)[\mu^{(n)}] = \lim_{s\to 0^+} \frac{\mathbb{E}_{\mu_0^n = \mu^{(n)}}(\Psi[\mu_s^{(n)}]) - \Psi(\mu^{(n)})}{s}$$

where

$$\mathbb{E}_{\mu_0^n = \mu^{(n)}}(\Psi[\mu_s^{(n)}]),$$

is the expectation of the functional $\Psi$ evaluated on the trajectory of the markov process $\mu_s^{(n)}$ taken on conditional on the initial step $\mu_0^{(n)} = \mu^{(n)}$.

1. Given our markov process i.e $\mu_t^{(n)}$ and $\mu_0^{(n)}$ we find the expression of the generator $\mathcal{L}_n\Psi[\mu^{(n)}]$ (using pertrubation analysis )
2. Since the process is markovian letting $t \to 0$ and considering the generator it will give us the evolution also between $t$ and $t + dt$ of $\Psi[\mu_t^{(n)}]$:

$$\partial_t \Psi(\mu_t^{(n)}) = (\mathcal{L}_n\Psi)[\mu_t^{(n)}], \Psi(\mu_t^{(n)})|_{t=0} = \Psi(\mu_0^{(n)})$$

3. Consider $n \to \infty$, identify the PDE corresponding to the generator

As $s \to 0$, and $\varepsilon \to 0$, we have:

$$E_0\Psi(q_s^n) - q^n = \underbrace{E_0\Psi(q_s^n) - E_0 q_{s-\varepsilon}^n}_{\text{weights updates}} + \underbrace{E_0 q_{s-\varepsilon}^n - q^n}_{\text{advection}}$$

The advection part:

$$\begin{aligned} A_n\Psi[q^n] &= \sum_{j=1}^n w_j \int \left\langle \nabla_x u_{p,q^{(n)}}(X^j)\delta_{X_j}(dx), \nabla_x D_{q^n}\Psi(X^j)\right\rangle \\ &= \int \langle \nabla_x u_{p,q^n}(x), \nabla_x D_{q^n}\Psi(x)\rangle) q^n(dx) \end{aligned}$$

For the weight update part note that we have:

$$w_s^j = w_{s-\varepsilon}^j + \varepsilon\alpha(u_{p,q_{s-\varepsilon}^n}(X_{s-\varepsilon}^j) - \mathbb{E}_{q_{s-\varepsilon}^{(n)}} u_{p,q_{s-\varepsilon}^n}) w_{s-\varepsilon}^j$$

$$q_s^n = \sum_{j=1}^N w_s^j \delta_{X_{s-\varepsilon}^j}$$

$$q_s^n = q_{s-\varepsilon}^n + \varepsilon'\alpha \sum_{j=1}^n w_{s-\varepsilon}^j (u_{p,q_{s-\varepsilon}^n}(X_{s-\varepsilon}^j) - \mathbb{E}_{q_{s-\varepsilon}^{(n)}} u_{p,q_{s-\varepsilon}^n}) \delta_{X_{s-\varepsilon}^j}$$

Hence we have:

$$\frac{q_s^n(x) - q_{s-\varepsilon}^n(x)}{\varepsilon'} = \alpha(u_{p,q_{s-\varepsilon}^n}(x) - \mathbb{E}_{q_{s-\varepsilon}^{(n)}} u_{p,q_{s-\varepsilon}^n}) q_{s-\varepsilon}^n(x) = \chi$$

Hence the variation of $\Phi$:

$$\lim_{\varepsilon' \to 0} \frac{\Psi(q_s^n) - \Psi(q_{s-\varepsilon}^n)}{\varepsilon'} = \int D_{q_{s-\varepsilon}^n}\Psi(x) d\chi(x) = \alpha \int D_{q_{s-\varepsilon}^n}\Psi(x)(u_{p,q_{s-\varepsilon}^n}(x) - \mathbb{E}_{q_{s-\varepsilon}^{(n)}} u_{p,q_{s-\varepsilon}^n}) q_{s-\varepsilon}^n(x) dx$$

As $s, \varepsilon \to 0$ we obtain the effect of weights updates as follows:

$$W_n\Psi[q^n] = \alpha \int D_{q^n}\Psi(x)(u_{p,q^n}(x) - \mathbb{E}_{q^{(n)}} u_{p,q^n}) q^n(x) dx$$

Hence the Generator has the following form:

$$(\mathcal{L}_n\Psi)[q^{(n)}] = \int \langle \nabla_x u_{p,q^n}(x), \nabla_x D_{q^n}\Psi(x)\rangle) q^n(dx) + \alpha \int D_{q^n}\Psi(x)(u_{p,q^n}(x) - \mathbb{E}_{q^{(n)}} u_{p,q^n}) q^n(x) dx$$

and we have:

$$\partial_t \Psi[q_t^n] = (\mathcal{L}_n\Psi)[q_t^{(n)}], with q_0^{(n)} = q$$

As $n \to \infty$ we have the evolution of the PDE:

$$\partial_t q_t = -div(q(x)\nabla_x u_{p,q_t}) + \alpha(u_{p,q_t} - \mathbb{E}_{q_t} u_{p,q_t})$$

and

$$\partial_t \Psi[q_t] = (\mathcal{L}\Psi)[q_t],$$

where $\mathcal{L}\Psi(q) = \int \langle \nabla_x u_{p,q}(x), \nabla_x D_q\Psi(x)\rangle\, q(dx) + \alpha \int D_q\Psi(x)(u_{p,q}(x) - \mathbb{E}_q u_{p,q}) q(x) dx.$

□

*Proof of Theorem 3 (Decrease of the MMD loss of the (Continous) Gradient Flow).* For $u_{p,q_t}^{\gamma,\lambda}$ we omit the up-scripts $\gamma$ and $\lambda$ in the following. Note that we have the following two expressions using the fact our functions are in the RKHS:

$$\begin{aligned} \int \langle \nabla_x u_{p,q_t}(x), \nabla_x \delta_{p,q_t}\rangle\, q_t(dx) &= \int \left\langle u_{p,q_t}, (J\Phi(x))^\top J\Phi(x)\delta_{p,q_t}\right\rangle q_t(dx) \\ &= \left\langle u_{p,q_t}, \mathbb{E}_{q_t}(J\Phi(x))^\top (J\Phi(x))\delta_{p,q_t}\right\rangle \\ &= \left\langle u_{p,q_t}, D(q_t)\delta_{p,q_t}\right\rangle. \end{aligned}$$

On the other hand:

$$
\begin{aligned}
& \int \delta_{p,q_t}(x)(u_{p,q_t}(x) - \gamma \mathbb{E}_{q_t} u_{p,q_t}) q_t(x) dx \\
= & \int \langle \delta_{p,q_t}, \Phi(x) \rangle \langle \Phi(x) - \gamma \mu(q_t), u_{p,q_t} \rangle q_t(x) dx \\
= & \int \langle \delta_{p,q_t}, \Phi(x) - \gamma \mu(q_t) \rangle \langle \Phi(x) - \gamma \mu(q_t), u_{p,q_t} \rangle q_t(x) dx \\
+ & \gamma \int \langle \delta_{p,q_t}, \mu(q_t) \rangle \langle \Phi(x) - \gamma \mu(q_t), u_{p,q_t} \rangle q_t(x) dx \\
= & \left\langle \delta_{p,q_t}, \left( \int (\Phi(x) - \gamma \mu(q_t)) \otimes (\Phi(x) - \gamma \mu(q_t)) q_t(dx) \right) u_{p,q_t} \right\rangle \\
+ & \gamma \langle \delta_{p,q_t}, \mu(q_t) \rangle \int \langle \Phi(x) - \gamma \mu(q_t), u_{p,q_t} \rangle q_t(x) dx \\
= & \langle \delta_{p,q_t}, C_\gamma(q_t) u_{p,q_t} \rangle + \underbrace{\gamma \langle \delta_{p,q}, \mu(q_t) \rangle \langle \mu(q_t) - \gamma \mu(q_t), u_{p,q_t} \rangle}_{=0, \text{ for } \gamma \in \{0,1\}} \\
= & \langle \delta_{p,q_t}, C_\gamma(q_t) u_{p,q_t} \rangle + 0.
\end{aligned}
$$

Consider $\Psi(q) = \frac{1}{2}\text{MMD}^2(p,q) = \frac{1}{2}\|\mu(p) - \mu(q)\|^2$, it is easy to see that the functional derivative wrt to $q$ is $D_q\Psi(q)(x) = -\delta_{p,q}$. Hence we have:

$$
\begin{aligned}
\frac{1}{2}\frac{d\text{MMD}^2(p,q_t)}{dt} &= -\int \langle \nabla_x u_{p,q_t}(x), \nabla_x \delta_{p,q_t} \rangle q_t(x) dx - \alpha \int \delta_{p,q_t}(x)(u_{p,q_t}(x) - \gamma \mathbb{E}_{q_t} u_{p,q_t}) q_t(x) dx \\
&= -\langle \delta_{p,q_t}, D(q_t) u_{p,q_t} \rangle - \alpha \langle \delta_{p,q_t}, C_\gamma(q_t) u_{p,q_t} \rangle \\
&= -\langle \delta_{p,q_t}, (D(q_t) + \alpha C_\gamma(q_t) + \lambda I - \lambda I) u_{p,q_t} \rangle \\
&= -(\langle \delta_{p,q_t}, (D(q_t) + \alpha C_\gamma(q_t) + \lambda I) u_{p,q_t} \rangle - \lambda \langle \delta_{p,q_t}, u_{p,q_t} \rangle) \\
&= -(\langle \delta_{p,q_t}, \delta_{p,q_t} \rangle - \lambda \langle \delta_{p,q_t}, u_{p,q_t} \rangle) \text{ where we used that } (D(q_t) + \alpha C_\gamma(q_t) + \lambda I) u_{p,q_t} = \delta_{p,q_t} \\
&= -\left(\text{MMD}^2(p,q_t) - \lambda \text{SF}^2_{\mathcal{H},\gamma,\lambda}(p,q_t)\right) \text{ by Definition of Sobolev-Fisher Distance} \\
&\leq 0
\end{aligned}
$$

since

$$
\text{MMD}^2(p,q_t) \geq \lambda \text{SF}^2_{\mathcal{H},\gamma,\lambda}(p,q_t)
$$

□

We now prove a Lemma the can be used to show that Unbalanced Sobolev descent has an acceleration advantage over Sobolev descent [2].

**Lemma 2.** *In the regularized case $\lambda > 0$ with $\alpha > 0$, the Kernel Sobolev-Fisher Discrepancy $SF_{\mathcal{H},\gamma,\lambda}$ is strictly upper bounded by the Kernel Sobolev discrepancy $\mathcal{S}_{\mathcal{H},\lambda}$ [2]:*

$$
SF^2_{\mathcal{H},\gamma,\lambda}(p,q) < \mathcal{S}^2_{\mathcal{H},\lambda}(p,q).
$$

*Proof.* Recall that (see Proposition 2):

$$
\text{SF}^2_{\mathcal{H},\gamma,\lambda}(p,q) = \left\langle (D + \alpha C_\gamma + \lambda I_m)^{-1} \delta_{p,q}, \delta_{p,q} \right\rangle,
$$

and that (see [2]):

$$
\mathcal{S}^2_{\mathcal{H},\lambda}(p,q) = \left\langle (D + \lambda I_m)^{-1} \delta_{p,q}, \delta_{p,q} \right\rangle.
$$

We now make use of the *Woodbury identity* $(A+B)^{-1} = A^{-1} - (A + AB^{-1}A)^{-1}$ with $A = D + \lambda I_m$ and $B = \alpha C_\gamma$, which allows us to write:

$$
(D + \alpha C_\gamma + \lambda I_m)^{-1} = (D + \lambda I_m)^{-1} - E, \tag{5}
$$

where $E = (A + AB^{-1}A)^{-1}$.

Notice that $A = D + \lambda I_m$ and $B = \alpha C_\gamma$ are both symmetric positive definite (SPD). Because the inverse of a SPD matrix is itself a SPD matrix, $B^{-1}$ is SPD. Because the product of SPD matrices is itself SPD, $AB^{-1}A$ is SPD. Because the inverse of the sum of SPD matrices is itself SPD, $E$ is SPD.

Equation (5) then implies:

$$(D + \alpha C_\gamma + \lambda I_m)^{-1} \prec (D + \lambda I_m)^{-1},$$

which, together with the definitions of $\mathrm{SF}^2_{\mathcal{H},\gamma,\lambda}$ and $\mathcal{S}^2_{\mathcal{H},\lambda}$, concludes the proof. □

# D Unbalanced Sobolev Descent With a Universal Kernel

While we presented the paper in a finite dimensional RKHS, to ease the presentation. We show in this Section, that our theory is general and apply to the infinite dimensional case. Of interest to us, is the case of a universal kernel. The convergence in MMD for a universal kernel implies the weak convergence in the distributional sense.

## D.1 Kernel Mean Embeddings, Covariance and Grammian of Derivatives Operators

Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space with an associated kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. We make the following assumptions on $\mathcal{H}$ as in [2]:

- A1 There exists $\kappa_1 < \infty$ such that $\sup_{x \in \mathcal{X}} \|k_x\|_{\mathcal{H}} < \kappa_1$.
- A2 The kernel is $C^2(\mathcal{X} \times \mathcal{X})$ and there exists $\kappa_2 < \infty$ such that for all $a = 1 \dots d$: $\sup_{x \in \mathcal{X}} Tr((\partial_a k)_x \otimes (\partial_a k)_x) < \kappa_2$.
- A3 $\mathcal{H}$ vanishes on the boundary (assuming $\mathcal{X} = \mathbb{R}^d$ it is enough to have for $f$ in $\mathcal{H}$ $\lim_{\|x\| \to \infty} f(x) = 0$).

The reproducing property give us that $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$ moreover $(D_a f)(x) = \frac{\partial}{\partial x_a} f(x) = \langle f, (\partial_a k)_x \rangle_{\mathcal{H}}$, where $(\partial_a k)_x(t) = \left\langle \frac{\partial k(s,.)}{\partial s_a}\Big|_{s=x}, k_t \right\rangle$. Note that those two quantities ($f(x)$ and $(D_a f)(x)$) are well defined and bounded thanks to assumptions A1 and A2 [28].
Similar to finite dimensional case we define the Gramian of derivatives operator of a distribution $q$ :

$$D(q) = \mathbb{E}_{x \sim \nu_q} \sum_{a=1}^{d} (\partial_a k)_x \otimes (\partial_a k)_x \; D(\nu_q) \in \mathcal{H} \otimes \mathcal{H} \tag{6}$$

The Kernel mean embedding is defined as follows:

$$\mu(p) = \mathbb{E}_{x \sim \nu_p} k_x \in \mathcal{H}. \tag{7}$$

The covariance operator is defined as follows for $\gamma \in \{0, 1\}$:

$$C_\gamma(q) = \mathbb{E}_{x \sim q} k_x \otimes k_x - \gamma \mu(q) \otimes \mu(q) \tag{8}$$

## D.2 Regularized Kernel Sobolev Fisher Discrepancy

Let $\lambda > 0, \alpha \geq 0$, similarly the Kernel Sobolev Fisher Discrepancy has the following form:

$$\mathrm{SF}^2_{\mathcal{H},\gamma,\lambda}(p, q) = \left\| (D(q) + \alpha C_\gamma(q) + \lambda I)^{-\frac{1}{2}} (\mu(\nu_p) - \mu(\nu_q)) \right\|^2_{\mathcal{H}},$$

where $D(q), \mu(q), C_\gamma(q)$ are defined in Equations (6),(7) and (8) respectively. The Sobolev Fisher witness function is defined as follows:

$$u^{\lambda,\gamma}_{p,q} = (D(q) + \alpha C_\gamma(q) + \lambda I)^{-1} (\mu(\nu_p) - \mu(\nu_q)) \in \mathcal{H}$$

its evaluation function is

$$u^{\lambda,\gamma}_{p,q}(x) = \left\langle (D(\nu_q) + \lambda I)^{-1} (\mu(\nu_p) - \mu(\nu_q)), k_x \right\rangle_{\mathcal{H}}$$

and its derivatives for $a = 1 \dots d$:

$$\partial_a u^{\lambda,\gamma}_{p,q}(x) = \left\langle (D(\nu_q) + \lambda I)^{-1} (\mu(\nu_p) - \mu(\nu_q)), \partial_a k_x \right\rangle_{\mathcal{H}}.$$

### D.3 USD with Infinite dimensional Kernel decreases the MMD distance

Theorem 3 holds for the infinite dimensional case. To see that it is enough to replace in the proof of Theroem 3 finite dimensional operators and embeddings $D(q), C_\gamma(q), \mu(q)$ with their infinite dimensional counterparts given in Equation in Equations (6),(7) and (8). All norms and dot products in $\mathbb{R}^m$, are also to be replaced with $\|.\|_{\mathcal{H}}$ and $\langle ., . \rangle_{\mathcal{H}}$.

## E Code and Hyper-parameters

Listing 1: Pytorch code for computing cost function $\mathcal{L}_S(\xi, \lambda)$ in Algorithm 3

```
import torch
from torch.autograd import grad

def descent_cost(f, x_p, w_p, x_q, w_q, lambda_aug, alpha, rho, gamma=1):
    """Computes the objective of Unbalance Sobolev Descent and returns the loss = -obj
    """
    x_q.requires_grad_(True)

    f_p, f_q = f(x_p), f(x_q)
    Ep_f = (w_p * f_p).mean()
    Eq_f = (w_q * f_q).mean()

    # FISHER
    constraint_F = (w_q * f_q**2).mean() - gamma * Eq_f**2

    # SOBOLEV
    grad_f_q = grad(outputs=Eq_f, inputs=x_q, create_graph=True)[0]
    normgrad_f2_q = (grad_f_q**2).sum(dim=1, keepdim=True)
    constraint_S = (w_q * normgrad_f2_q).mean()

    # Combining FISHER and SOBOLEV constraints
    constraint_tot = (constraint_S + alpha * constraint_F - 1.0)

    obj_f = Ep_f - Eq_f \
           - lambda_aug * constraint_tot - rho/2 * constraint_tot**2

    return -obj_f, Ep_f, Eq_f, normgrad_f2_q
```

## F Architecture of Neural Network discriminator

```
D_mlp = Sequential(
    (L0): Linear(in_features=n_inputs, out_features=n_layers[0], bias=True)
    (N0): ReLU(inplace=True)
    (L1): Linear(in_features=n_layers[0], out_features=n_layers[1], bias=True)
    (N1): ReLU(inplace=True)
    (D1): Dropout(p=0.2, inplace=False)
    (L2): Linear(in_features=n_layers[1], out_features=n_layers[2], bias=True)
    (N2): ReLU(inplace=True)
    (V): Linear(in_features=n_layers[2], out_features=1, bias=False)
)
```

## G Hyperparameters for experiments

Listing 2: Hyperparameters for synthetic experiments (Figs. 1, 2, 5, 6)

```
{
    "n_layers": [64, 1024, 64],  # Number of neurons in hidden layers of discriminator
    "n_points_src": 4000,        # Number of points sampled from source distribution
    "n_points_target": 4000,     # Number of points sampled from target distribution
    "T": 800,                    # Number descent steps
    "optimizer": Adam(amsgrad=True) # Optimizer for discriminator (reset at every update of distribution q)
    "batchSize": 512,            # Batch size for discriminator updates
    "n_c_startup": 200,          # Number of steps for discriminator updates at startup
    "n_c": 20,                   # Number of steps for discriminator updates in-between updates of distribution q
    "wdecay": 1e-5,              # Weight decay factor
    "lrD": 1e-4,                 # Learning rate for discriminator updates
    "lrQ": 1e-4,                 # Learning rate for updates of distribution q
    "tau": 1e-3,                 # Birth-death rate
    "alpha": 0.6,                # Damping factor ($\alpha$ in Algorithm 3)
    "lambda_aug_init": 1e-5,     # Initialization of augmented Lagrange multiplier (in Algorithm 3)
    "rho": 1e-6                  # Learning rate of augmented Lagrange multiplier
}
```

Listing 3: Hyperparameters for color transfer experiments (Figs. 3, 7)

```
{
    "n_layers": [128, 2048, 128], # Number of neurons in hidden layers of discriminator
    "n_points_src": 65536,        # Number of points sampled from source distribution
    "n_points_target": 65536,     # Number of points sampled from target distribution
    "T": 800,                     # Number descent steps
    "optimizer": Adam(amsgrad=True) # Optimizer for discriminator (reset at every update of distribution q)
    "batchSize": 500,             # Batch size for discriminator updates
    "n_c_startup": 300,           # Number of steps for discriminator updates at startup
    "n_c": 5,                    # Number of steps for discriminator updates in-between updates of distribution q
    "wdecay": 1e-5,               # Weight decay factor
    "lrD": 1e-4,                  # Learning rate for discriminator updates
    "lrQ": 1e-4,                  # Learning rate for updates of distribution q
    "tau": 1e-6,                  # Birth-death rate
    "alpha": 0.3,                 # Damping factor ($\alpha$ in Algorithm 3)
    "lambda_aug_init": 0.0,       # Initialization of augmented Lagrange multiplier (in Algorithm 3)
    "rho": 1e-6                   # Learning rate of augmented Lagrange multiplier
}
```

Listing 4: Hyperparameters for single-cell analysis interpolation experiments (Fig. 4, 8)

```
{
    "n_layers": [128, 1024, 64], # Number of neurons in hidden layers of discriminator
    "n_points_src": 3500,        # Number of points sampled from source distribution
    "n_points_target": 3500,     # Number of points sampled from target distribution
    "T": 400,                    # Number descent steps
    "optimizer": Adam(amsgrad=True) # Optimizer for discriminator (reset at every update of distribution q)
    "batchSize": 100,            # Batch size for discriminator updates
    "n_c_startup": 300,          # Number of steps for discriminator updates at startup
    "n_c": 5,                   # Number of steps for discriminator updates in-between updates of distribution q
    "wdecay": 1e-5,              # Weight decay factor
    "lrD": 1e-4,                 # Learning rate for discriminator updates
    "lrQ": 1e-4,                 # Learning rate for updates of distribution q
    "tau": 2e-4,                 # Birth-death rate
    "alpha": 0.2,                # Damping factor ($\alpha$ in Algorithm 3)
    "lambda_aug_init": 1e-5,     # Initialization of augmented Lagrange multiplier (in Algorithm 3)
    "rho": 1e-6                  # Learning rate of augmented Lagrange multiplier
    "normalization": nn.BatchNorm1d(track_running_stats=False, momentum=0.0) # Substitutes dropout layer after second hidden
          layer
}
```

# H Additional Plots

## H.1 Synthetic Examples

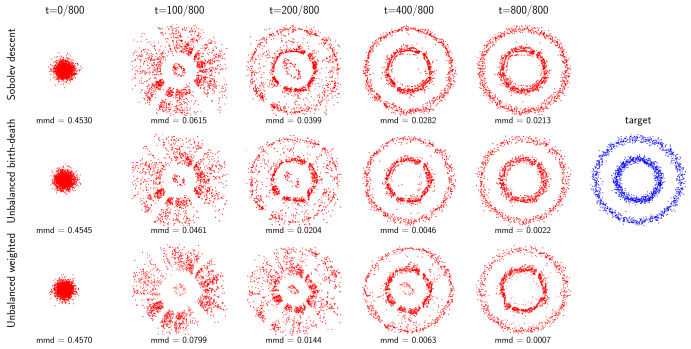
We give in Figs 5 and 6 additional synthetic experiments:

## H.2 Image Coloring

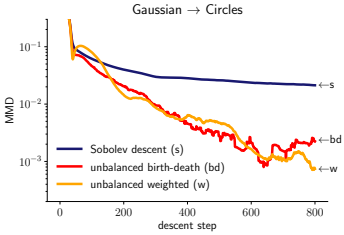
We give in Figure 7, the trajectories of the descent in image color transfer experiment.

## H.3 Comparisons to Waddington Optimal Transport for single-cell analysis

We give in Figure 8 the evolution of the MMD as function of the day of interpolation using USD and unbalanced OT as in WOT.

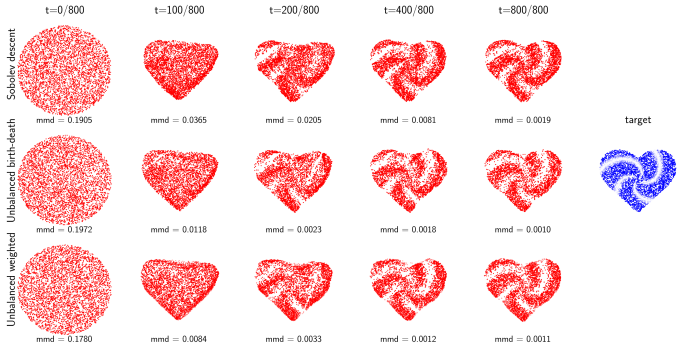



(a) Neural Unbalanced Sobolev Descent paths in transporting a Gaussian to circles). We compare Sobolev descent (SD, [2]) to both USD implementations with birth and death processes (bd: Algorithm 2) as well as the weighted version implementation (w: Algoritm 1, note that in this case we overlay the points with their respective weights where coloring density encodes the weights). We see that birth and death processes helps USD to outperform SD in capturing the two modes.

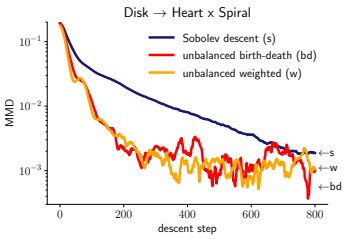



(b) MMD function of the time in the descent from a Gaussian to Circles: We see that birth and death processes in both implementations of USD accelerate the convergence to the target distribution and reaches lower MMD than Sobolev Descent that relies on advection only.

Figure 5: Neural Unbalanced Sobolev Descent transporting a Gaussian to circles (target samples have uniform weights, $a_j = \frac{1}{n}$).




(a) Neural Unbalanced Sobolev Descent paths in transporting a disk to a heart/spiral. We compare Sobolev descent (SD, [2]) to both USD implementations with birth and death processes (bd: Algorithm 2) as well as the weighted version implementation (w: Algoritm 1, note that in this case we overlay the points with their respective weights where coloring density encodes the weights). We see that birth and death processes helps USD to outperform SD in capturing the two modes.




(b) MMD function of the time in the descent from a disk to a heart/spiral: We see that birth and death processes in both implementations of USD accelerate the convergence to the target distribution and reaches lower MMD than Sobolev Descent that relies on advection only.

Figure 6: Neural Unbalanced Sobolev Descent transporting a 'disk' to a 'heart' weighted by a spiral-shaped gradient.


Figure 7: Color Transfer with USD using (bd) Algorithm 2. Trajectories of the descent.

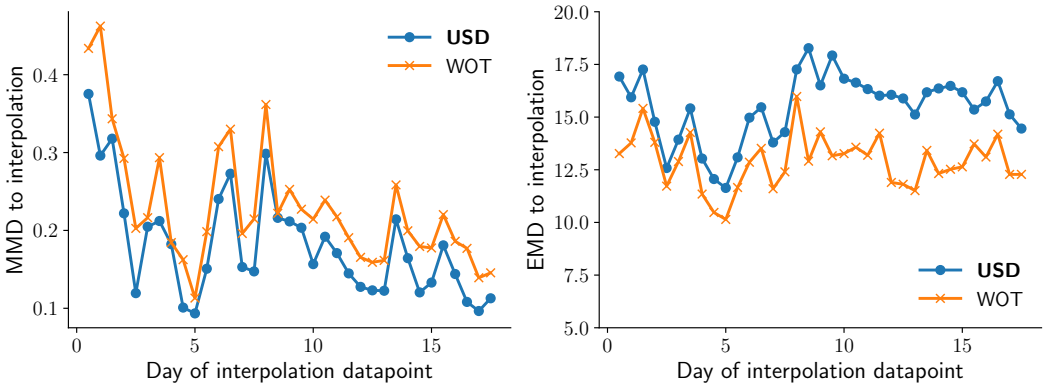



Figure 8: MMD and EMD between predicted mid points (using USD and WOT) and their respective ground truths as function of the day of interpolation.