IBM Research Report

Regularized Finite Dimensional Kernel Sobolev Discrepancy

Youssef Mroueh
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Regularized Finite Dimensional Kernel Sobolev Discrepancy

Youssef Mroueh
IBM Research
IBM T.J Watson Research Center
mroueh@us.ibm.com

Abstract

We show in this note that the Sobolev Discrepancy introduced in [1] in the context of generative adversarial networks, is actually the weighted negative Sobolev norm $\|.\|_{\dot{H}^{-1}(\nu_q)}$, that is known to linearize the Wasserstein W_2 distance and plays a fundamental role in the dynamic formulation of optimal transport of Benamou and Brenier. Given a Kernel with *finite dimensional feature map* we show that the Sobolev discrepancy can be approximated from finite samples. Assuming this discrepancy is finite, the error depends on the approximation error in the function space induced by the finite dimensional feature space kernel and on a statistical error due to the finite sample approximation.

1   Sobolev Discrepancy and Weighted Negative Sobolev Norms

In this Section we review the Sobolev Discrepancy introduced in [1]. Let $\mathcal{X}$ be a compact space in $\mathbb{R}^d$ with lipchitz boundary $\partial\mathcal{X}$. We start by defining the Sobolev Discrepancy:

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

$$D : \mathcal{S}(\nu_p, \nu_q) = \sup_f \left\{ \mathbb{E}_{x\sim\nu_p} f(x) - \mathbb{E}_{x\sim\nu_q} f(x) : f \in W_0^{1,2}(\mathcal{X}, \nu_q), \mathbb{E}_{x\sim\nu_q} \|\nabla_x f(x)\|^2 \leq 1 \right\} \quad (1)$$

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

We note here that this Sobolev discrepancy is actually known and already studied in optimal transport and relates to the Wasserstein 2 distance and its dynamical form given by Benamou and Brenier [2]. It is indeed defined through the weighted negative Sobolev Norm :

Definition 2 (Weighted Negative Sobolev Norm [3, 4])**.** *For μ a positive measure on $\mathcal{X}$, For a signed measure χ on $\mathcal{X}$, the weighted negative Sobolev Norm is defined as follows:*

$$\|\chi\|_{\dot{H}^{-1}(\mu)} = \sup_{f, \int_{\mathcal{X}} \|\nabla_x f(x)\|^2 d\mu(x) \leq 1} \left| \int_{\mathcal{X}} f(x) d\chi(x) \right|.$$

$\|\chi\|_{\dot{H}^{-1}(\mu)}$ is the dual norm of the weighted Sobolev semi-norm $\|f\|_{\dot{H}(\mu)} = \int_{\mathcal{X}} \|\nabla_x f(x)\|^2 d\mu(x)$. This norm is finite for measures of zero total mass, and can be infinite.

It follows therefore that the Sobolev discrepancy corresponds to the $\dot{H}^{-1}(\nu_q)$ norm:

$$\mathcal{S}(\nu_p, \nu_q) = \|\nu_p - \nu_q\|_{\dot{H}^{-1}(\nu_q)}. \quad (2)$$

Note that $\dot{H}^{-1}(\nu_q)$ is different from the negative Sobolev norm $\dot{H}^{-1}(\mathcal{X})$ defined as follows for a signed measure χ (See for instance [5] chapter 5 Section 5.5.2)

$$\|\chi\|_{\dot{H}^{-1}(\mathcal{X})} = \sup_{f, \int_{\mathcal{X}} \|\nabla_x f(x)\|^2 dx \leq 1} \left| \int_{\mathcal{X}} f(x) d\chi(x) \right|$$

Note the negative Sobolev norm defines the following distance between ν_p and ν_q:

$$\|\nu_p - \nu_q\|_{\dot{H}^{-1}(\mathcal{X})} = \sup_{f, \int_{\mathcal{X}} \|\nabla_x f(x)\|^2 dx \leq 1} \int_{\mathcal{X}} f(x) d(\nu_p(x) - \nu_q(x)). \tag{3}$$

Note that negative Sobolev norms relate to Energy distances and MMD (See for instance [6] page 119)

2 Dual and Primal Formulations of the Sobolev Discrepancy

Following [5] (Chapter 5 Section 5.5.2) that characterizes the solution of problem 3 via a diffusion PDE, we characterize in this Section the solution of Problem (1) via an advection PDE. This connection to advection PDE was already given in [1] and is actually also a known result in the optimal transport literature. We simplify here the proofs and notations, and give a primal $\inf$ formulation for the $\sup$ "dual" formulation in (1).

Proposition 1 (Primal and Dual Formulation of Sobolev Discrepancy). *The following holds true:*

*1) **Dual witness function from an advection PDE:** Let $u_{p,q}$ be the solution of the following advection PDE:*
$$p(x) - q(x) = -div(q(x)\nabla_x u(x)) \ \ u|_{\partial\mathcal{X}} = 0, \tag{4}$$
physically this PDE means that we are moving the mass q to p following the flows of a velocity field given by $\nabla_x u$. We have $\mathcal{S}^2(\nu_p, \nu_q) = \int_{\mathcal{X}} \|\nabla_x u_{p,q}\|^2 q(x)dx$, and the witness function of the Dual formulation of (1) is given by: $f^*_{\nu_p, \nu_q}(x) = \frac{u_{p,q}(x)}{\mathcal{S}(\nu_p, \nu_q)}$.

*2) **Sobolev Discrepancy as a minimum kinetic energy under an advection transport of ν_q to ν_p:** $\mathcal{S}(\nu_p, \nu_q)$ given in (1) admits the following equivalent primal formulation:*

$$P : \mathcal{S}(\nu_p, \nu_q) = \inf_{v: \mathcal{X} \to \mathbb{R}^d} \left\{ \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} \text{ subject to} : p(x) - q(x) = -div(q(x)v(x)) \right\} \tag{5}$$

that is the minimum kinetic energy to advect the mass q to p following the velocity field given by v. The optimal velocity field is given by $v^* = \nabla_x u_{p,q}$, where $u_{p,q}$ is the solution of the advection PDE (4).

It is important to note that the primal formulation (17) gives the transport interpretation of the Sobolev Discrepancy . $\mathcal{S}^2(\nu_p, \nu_q)$ is the minimum kinetic energy to transport the mass from the source measure ν_q to the target measure ν_p via an advection equation (4) with a velocity field $v = \nabla_x u$. The kinetic energy is measured with the velocity field v and the mass of the source distribution q. The optimal velocity field is a gradient $v^*(x) = \nabla_x u_{p,q}(x)$, and $\mathcal{S}^2(\nu_p, \nu_q) = \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x)dx$. The dual form (1) is computationally friendly since both its objective and constraints can be expressed as expectations on ν_p and ν_q, and can be seen as a regularized mean discrepancy. Those computational properties were exploited in [1].

Remark 1. *As a summary of this section the Sobolev discrepancy admits three equivalent interpretations: 1) A "Computational friendly" regularized mean discrepancy in its dual form given in (1). 2) A minimum Kinetic Energy in a transport via advection in its primal given in (17). 3) As a discrepancy computed with weighted Negative Sobolev Norms $\|.\|_{\dot{H}^{-1}(\nu_q)}$ given in Equation (2).*

3 Sobolev Discrepancy and the Wasserstein 2 Distance

3.1 Wasserstein 2: Static and Dynamic Formulation

Given the transport via advection interpretation of the Sobolev discrepancy, we think in the following of ν_q as the source distribution and ν_p as the target distribution. The relation between weighted

negative Sobolev norms and the Wasserstein 2 distance is well established in the optimal transport literature, since it linearizes the Wasserstein 2 distance [3]. The Wasserstein 2 distance is defined as follows:

$$W_2(\nu_q, \nu_p) = \left\{ \min_{(X,Y)} \mathbb{E}_{(X,Y)} \|X - Y\|_2^2, \ X \sim \nu_q, Y \sim \nu_p \right\}$$

For a small perturbation χ and any measure μ and a small ε (See for instance [3]):

$$W_2(\mu, \mu + \varepsilon\chi) = \varepsilon \|\chi\|_{\dot{H}^{-1}(\mu)} + o(\varepsilon).$$

This identity is at the heart of the dynamic formulation of optimal transport [2]:

$$W_2(\nu_q, \nu_p) = \int_0^1 \|d\mu_{q_t}\|_{\dot{H}^{-1}(\mu_{q_t})}, \mu_0 = q, \mu_1 = p.$$

The dynamic formulation as given by Benamou and Brenier [2], finds a path of densities for transporting q to p via advection while minimizing the kinetic energy $\|d\mu_{q_t}\|_{\dot{H}^{-1}(\mu_{q_t})}$. This can be written in the following equivalent form. For $t \in [0,1]$, let $v_t : \mathcal{X} \to \mathbb{R}^d$ be velocity fields and μ_{q_t} be intermediate measures whose densities are q_t, we have:

$$W_2^2(\nu_q, \nu_p) = \min_{q_t, \boldsymbol{v}_t} \left\{ \int_0^1 \int_{\mathcal{X}} \|\boldsymbol{v}_t(x)\|^2 q_t(x) dx dt, \ \frac{\partial q_t}{\partial t} = -div(q_t \boldsymbol{v}_t), q_{t=0} = q, q_{t=1} = p. \right\} \quad (6)$$

Note that the expression given in Equation (6) is exploiting the primal kinetic energy formulation of the *Sobolev discrepancy* given in Equation (17). Peyre [4] exploited this connection between the Wasserstein distance and the the weighted negative Sobolev norm, to give upper and lower bounds on W_2 and $\|.\|_{\dot{H}^{-1}(\nu_q)}$. In the following, we give upper and lower bounds on W_2 and the Sobolev Discrepancy $\mathcal{S}(\nu_p, \nu_q)$, while imposing stronger assumption on the boundedness of the density as done in [5] (Chapter 5, Section 5.5.2). Note that [5] gives upper and lower bounds for Negative Sobolev norms $\|.\|_{\dot{H}^{-1}(\mathcal{X})}$ and not for the weighted case $\|.\|_{\dot{H}^{-1}(\nu_q)}$ as done in [4].

3.2 Bounding W_2 with Sobolev Discrepancy

The following proposition shows that under some regularity conditions the Wasserstein 2 distance can be upper and lower bounded by the Sobolev Discrepancy.

Proposition 2. *Assume that ν_p, ν_q are absolutely continuous measures, with densities bounded from above and below by two constants $(0 < a < b < \infty)$. Then we have:*

$$\sqrt{\frac{a}{b}} \mathcal{S}(\nu_p, \nu_q) \leq W_2(\nu_q, \nu_p) \leq 2\mathcal{S}(\nu_p, \nu_q).$$

From Proposition 2 we see that the Wasserstein 2 distance W_2 and the Sobolev Discrepancy are equivalent under some regularity assumptions on the density.

3.3 Unconstrained Form of $\mathcal{S}^2(\nu_p, \nu_q)$

We end this Section with an unconstrained equivalent form for the Sobolev discrepancy that will prove to be useful for the our future developments in the paper:

Lemma 1. *The following equivalent form holds true for the squared Sobolev Discrepancy:*

$$\mathcal{S}^2(\nu_p, \nu_q) = \sup_{u \in W_0^{1,2}(\mathcal{X}, \nu_q)} \left\{ L(u) = 2 \int_{\mathcal{X}} u(x) d(\nu_p(x) - \nu_q(x)) - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 d\nu_q(x) \right\},$$

$$\tag{7}$$

the optimal u^* is given by $u_{p,q}$ solution of the advection PDE (4). Moreover we have for any feasible u:

$$\mathcal{S}^2(\nu_p, \nu_q) - L(u) = \int_{\mathcal{X}} \|\nabla_x u(x) - \nabla_x u_{p,q}(x)\|^2 q(x) dx. \quad (8)$$

Given in this form we see that the main computational difficulty in computing the Sobolev Discrepancy is in optimization over the space $W_0^{1,2}(\mathcal{X}, \nu_q)$. [1] proposed to parametrize functions with neural networks. In this paper we propose to relax this function space to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, with the goal of having certain of the nice theoretical propreties of the Sobolev Discrepancies carrying on to the Kernelized case.

3

4 Kernelized Sobolev Discrepancy

In this Section we define the Kernelized Sobolev Discrepancy by looking for the optimal witness funtion of (7) in a Hypothesis function class that is a **Finite dimensional** Reproducing Kernel Hilbert Space (RKHS). We start first by reviewing some RKHS properties and assumptions needed for our statements.

Let $\mathscr{H}$ be a Reproducing Kernel Hilbert Space with an associated finite feature map $\Phi : \mathcal{X} \to \mathbb{R}^m$. The associated kernel k is therefore $k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}} = \sum_{j=1}^{m} \Phi_j(x)\Phi_j(y)$. Note that for any function $f \in \mathscr{H}$, we have $f(x) = \langle \boldsymbol{f}, \Phi(x) \rangle$, where $\boldsymbol{f} \in \mathbb{R}^m$, and $\langle, \rangle$ is the dot product in $\mathbb{R}^m$. We note $\|f\|_{\mathscr{H}} = \|\boldsymbol{f}\| = \sum_{j=1}^{m} \boldsymbol{f}_j^2$. 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 $\nabla_x f(x) = (J\Phi(x)\boldsymbol{f}) \in \mathbb{R}^m$.

We make the following assumptions on $\mathscr{H}$:

A1 There exists $\kappa_1 < \infty$ such that $\sup_{x \in \mathcal{X}} \|\Phi(x)\| < \kappa_1$.

A2 There exists $\kappa_2 < \infty$ such that for all $a = 1 \dots d$:
$\sup_{x \in \mathcal{X}} Tr(\frac{\partial}{\partial x_a}\Phi(x) \otimes \frac{\partial}{\partial x_a}\Phi(x)) < \kappa_2$.

A3 $\mathscr{H}$ vanishes on the boundary: $\forall j = 1 \dots m, \Phi_j(x)|_{\partial\mathcal{X}} = 0$ (for $\mathcal{X} = \mathbb{R}^d$, it is enough to have $\lim_{x \to \infty} \Phi_j(x) = 0$) .

4.1 Kernel Sobolev Discrepancy

We define in what follows the Kernelized Sobolev Discrepancy by restricting the problem given in Equation to (1) to functions in a RKHS.

Definition 3 (Kernelized Sobolev Discrepancy). *Let $\mathscr{H}$ be a finite dimensional RKHS satisfying assumptions A1,A2 and A3. Let ν_p, ν_q be two measures defined on $\mathcal{X}$. We define the Sobolev discrepancy restricted to the space $\mathscr{H}$ as follows:*

$$\mathcal{S}_{\mathscr{H}}(\nu_p, \nu_q) = \sup_f \left\{ \mathbb{E}_{x \sim \nu_p} f(x) - \mathbb{E}_{x \sim \nu_q} f(x), f \in \mathscr{H}, \mathbb{E}_{x \sim \nu_q} \|\nabla_x f(x)\|^2 \leq 1 \right\} \qquad (9)$$

we note $f_{\nu_p,\nu_q}^{\mathscr{H}} \in \mathscr{H}$, the optimal witness function.

Note $\Omega(f) = \mathbb{E}_{x \sim \nu_q} \|\nabla_x f(x)\|^2$. For $f \in \mathscr{H}$, we have:

$$\Omega(f) = \sum_{a=1}^{d} \mathbb{E}_{x \sim \nu_q} \left\langle \boldsymbol{f}, \frac{\partial \Phi(x)}{\partial x_a} \right\rangle^2 = \sum_{a=1}^{d} \mathbb{E}_{x \sim \nu_q} \left\langle \boldsymbol{f}, \left(\frac{\partial \Phi(x)}{\partial x_a} \otimes \frac{\partial \Phi(x)}{\partial x_a} \right) \boldsymbol{f} \right\rangle = \langle f, D(\nu_q))f \rangle_{\mathscr{H}},$$

where we identified an operator $D(\nu_q)$:

$$D(\nu_q) = \mathbb{E}_{x \sim \nu_q} \sum_{a=1}^{d} \frac{\partial \Phi(x)}{\partial x_a} \otimes \frac{\partial \Phi(x)}{\partial x_a} = \mathbb{E}_{x \sim \nu_q}[J\Phi(x)]^\top J\Phi(x). \qquad (10)$$

We call $D(\nu_q)$ the Kernel Derivative Gramian Embedding **KDGE** of a distribution ν_q . KDGE is an operator embedding of the distribution i, that takes the fingerprint of the distribution with respect to the feature map derivatives averaged over all coordinates. This operator embedding of ν_q is to be contrasted with the classic Kernel Mean Embedding **KME** of a distribution in $\mathscr{H}$: $\mu(\nu_q) = \mathbb{E}_{x \sim \nu_q} \Phi(x)$.

Lemma 2 (Unconstrained Form of Kernel Sobolev Discrepancy).

$$\begin{aligned}
\mathcal{S}_{\mathscr{H}}^2(\nu_p, \nu_q) &= \sup_{u \in \mathscr{H}} \left\{ 2 \int_{\mathcal{X}} u(x)d(\nu_p(x) - \nu_q(x)) - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 d\nu_q(x) \right\} \\
&= \sup_{\boldsymbol{u} \in \mathbb{R}^m} 2 \langle \boldsymbol{u}, \mu(\nu_p) - \mu(\nu_q) \rangle - \langle \boldsymbol{u}, (D(\nu_q))\boldsymbol{u} \rangle, \qquad (11)
\end{aligned}$$

where $\mu(\nu_p), D(\nu_q)$ are the KME the KDGE defined above. Let $\boldsymbol{u}^* = \boldsymbol{u}_{p,q}^{\mathscr{H}}$, be the optimum.

Proof. The proof follows from Proposition 1,setting $\lambda = 0$. $\qquad \square$

Proposition (3) gives the expression of the optimal Kernel Sobolev witness function $f_{\nu_p,\nu_q}^{\mathscr{H}} \in \mathscr{H}$.

Proposition 3. *Assume that the KDGE of ν_q, $D(\nu_q)$ defined in Equation* (10) *is non singular. Let $\mu(\nu_p)$ and $\mu(\nu_q)$ be the KME of ν_p and ν_q respectively. The solution of Problem* (11), $\boldsymbol{u}_{p,q}^{\mathscr{H}}$ *in $\mathbb{R}^m$ is given by:*

$$\boldsymbol{u}_{p,q}^{\mathscr{H}} = [D(\nu_q)]^{-1}\left(\mu(\nu_p) - \mu(\nu_q)\right), \tag{12}$$

and $u_{p,q}^{\mathscr{H}} = \langle \boldsymbol{u}_{p,q}^{\mathscr{H}}, \Phi(x)\rangle.$ $\mathcal{S}_{\mathscr{H}}^2(\nu_p,\nu_q) = \left\| [D(\nu_q)]^{-\frac{1}{2}}\left(\mu(\nu_p) - \mu(\nu_q)\right)\right\|^2 =$ $\int_{\mathcal{X}} \left\| \nabla_x u_{p,q}^{\mathscr{H}}(x)\right\|^2 q(x)dx,$ *and the witness function of Kernel Sobolev Discrepancy* (9) $f_{\nu_p,\nu_q}^{\mathscr{H}} = \frac{\boldsymbol{u}_{\nu_p,\nu_q}^{\mathscr{H}}}{\mathcal{S}_{\mathscr{H}}(\nu_p,\nu_q)}$

Note that $u_{p,q}^{\mathscr{H}}$ is the approximation in $\mathscr{H}$ of $u_{p,q}$ the solution of the PDE (4). We have from Lemma 1 (Equation (8)):

$$\mathcal{S}^2(\nu_p,\nu_q) - \mathcal{S}_{\mathscr{H}}^2(\nu_p,\nu_q) = \int_{\mathcal{X}} \left\| \nabla_x u_{p,q}^{\mathscr{H}}(x) - \nabla_x u_{p,q}(x)\right\|^2 q(x)dx,$$

Hence the approximation in the space $\mathscr{H}$ is in the Sobolev semi-norm sense. We know that $\nabla_x u_{p,q}$ has the physical interpretation of a velocity advecting the mass from q to p. In the next Section we will take a close look at $\nabla_x u_{p,q}^{\mathscr{H}}$.

4.2 Transport in RKHS: Understanding $D(\nu_q)$

Let $(\lambda_j, \boldsymbol{\psi}_j)$ be eigenvectors of $D(\nu_q)$. Assume that $\lambda_j > 0$ for all $j = 1\ldots m$. We have : $D(\nu_q)\boldsymbol{\psi}_j = \lambda_j\boldsymbol{\psi}_j$, this means that $\left\langle \frac{1}{\sqrt{\lambda_k}}\boldsymbol{\psi}_k, D(\nu_q)\frac{1}{\sqrt{\lambda_j}}\boldsymbol{\psi}_j\right\rangle = \delta_{jk}$. Note $\tilde{\boldsymbol{\psi}}_j = \frac{1}{\sqrt{\lambda_j}}\boldsymbol{\psi}_j$. It is easy to see that this means: $\left\langle \frac{1}{\sqrt{\lambda_k}}\boldsymbol{\psi}_k, D(\nu_q)\frac{1}{\sqrt{\lambda_j}}\boldsymbol{\psi}_j\right\rangle = \int_{\mathcal{X}} \left\langle \nabla_x\tilde{\psi}_j(x), \nabla_x\tilde{\psi}_k(x)\right\rangle_{\mathbb{R}^d} q(x)dx = \delta_{jk}$, hence we have $\{\nabla_x\tilde{\psi}_j\}_{j=1}^m$ are orthonormal in $\mathscr{L}_2(\mathcal{X},\nu_q)^{\otimes d}$. We have for all $a = 1\ldots d$:

$$\frac{\partial u_{p,q}^{\mathscr{H}}(x)}{\partial x_a} = \left\langle \boldsymbol{u}_{\nu_p,\nu_q}^{\mathscr{H}}, \frac{\partial\Phi(x)}{\partial x_a}\right\rangle = \left\langle [D(\nu_q)]^{-1}\left(\mu(\nu_p) - \mu(\nu_q)\right), \frac{\partial\Phi(x)}{\partial x_a}\right\rangle$$

$$= \left\langle \sum_{j=1}^m \frac{1}{\lambda_j}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^*(\mu(\nu_p) - \mu(\nu_q)), \frac{\partial\Phi(x)}{\partial x_a}\right\rangle = \sum_{j=1}^m \frac{1}{\lambda_j}\langle\boldsymbol{\psi}_j, \mu(\nu_p) - \mu(\nu_q)\rangle\left\langle\boldsymbol{\psi}_j, \frac{\partial\Phi(x)}{\partial x_a}\right\rangle$$

$$= \sum_{j=1}^m \frac{1}{\lambda_j}\langle\boldsymbol{\psi}_j, \mu(\nu_p) - \mu(\nu_q)\rangle\frac{\partial\psi_j(x)}{\partial x_a}.$$

Hence we have: $\nabla_x u_{p,q}^{\mathscr{H}}(x) = \sum_{j=1}^m \left\langle\tilde{\psi}_j, \mu(\nu_p) - \mu(\nu_q)\right\rangle\nabla_x\tilde{\psi}_j(x)$, hence $D(\nu_q)$ eigenvectors give raise to a basis of principal transport directions $\nabla_x\tilde{\psi}_j(x) = J\Phi(x)\tilde{\boldsymbol{\psi}}_j$. A principal transport direction is weighted positively if $\left\langle\tilde{\psi}_j, \mu(\nu_p) - \mu(\nu_q)\right\rangle > 0$, meaning it is aligned with the KME differences in the direction of the desired transport from q to p.

4.3 Kernel Sobolev Discrepancy and the Wasserstein 2 Distance

In this Section we show that if $W_2(\nu_q,\nu_p) = 0$ this implies that the finite dimensional Kernel Sobolev Discrepancy is zero. Which means that a sequence is convergent in the Kernel Sobolev Discrepancy whenever it converges in the W_2 sense.

Proposition 4 (Convergence and Density). *Assume ν_p and ν_q are continuous and bounded from above an below by $0 < a < b$ For a RKHS $\mathscr{H}$ with finite dimensional feature map satisfying Assumptions A1, A2 and A3. We have: $\mathcal{S}_{\mathscr{H}}(\nu_p,\nu_q) \leq \mathcal{S}(\nu_p,\nu_q)$, and :*

$$\sqrt{\frac{a}{b}}\mathcal{S}_{\mathscr{H}}(\nu_p,\nu_q) \leq W_2(\nu_q,\nu_p).$$

which means that a sequence ν_{q_n} (continuous with densities, bounded from above and below) is convergent in $\mathcal{S}_{\mathscr{H}}$, whenever it converges in the Wasserstein 2 W_2.

5 Regularized Kernel Sobolev Discrepancy

Regularization in the RKHS consists as we will see in avoiding singularity issues of the KDGE $D(\nu_q)$ and plays a fundamental role in stabilizing the computations of the Discrepancy. We define below the Regularized Kernel Sobolev Discrepancy (RKSD):

Definition 4 (Regularized Kernel Sobolev Discrepancy (RKSD)). *The RKSD is defined as follows,*

$$\mathcal{S}_{\mathscr{H},\lambda}(\nu_p,\nu_q) = \sup_f \left\{ \mathbb{E}_{x\sim\nu_p} f(x) - \mathbb{E}_{x\sim\nu_q} f(x), f \in \mathscr{H}, \mathbb{E}_{x\sim\nu_q} \|\nabla_x f(x)\|^2 + \lambda \|f\|_{\mathscr{H}}^2 \leq 1 \right\}, \tag{13}$$

where $\lambda > 0$ is the regularization parameter and $\|.\|_{\mathscr{H}}$ is the RKHS norm. Let $f_{\nu_p,\nu_q}^{\lambda}$ be the optimal witness function.

The following proposition summarizes the main properties of RKSD:

Theorem 1 (Unconstrained Form of the RKSD/ witness function). *The squared RKSD has the following equivalent form.*

$$\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q) = \sup_{u\in\mathscr{H}} \left\{ 2\int_{\mathcal{X}} u(x)d(\nu_p(x) - \nu_q(x)) - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 d\nu_q(x) - \lambda \|u\|_{\mathscr{H}}^2 \right\}$$
$$= \sup_{u\in\mathbb{R}^m} L(u,\lambda) = 2\langle \boldsymbol{u}, \mu(\nu_p) - \mu(\nu_q)\rangle - \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u}\rangle \tag{14}$$

The following properties characterize the RKSD and its witness function

1) The optimal u^* in (14) is $\boldsymbol{u}_{p,q}^{\lambda} = (D(\nu_q) + \lambda I)^{-1}(\mu(\nu_p) - \mu(\nu_q))$.

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

3) $\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q) = \int_{\mathcal{X}} \|\nabla_x u_{p,q}^{\lambda}(x)\|^2 q(x)dx + \lambda \|\boldsymbol{u}_{p,q}^{\lambda}\|^2$, hence $\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q)$ is a regularized kinetic energy.

4) For any $u \in \mathscr{H}$ we have:

$$\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q) - L(u,\lambda) = \left\| \sqrt{D(\nu(q))}(\boldsymbol{u} - \boldsymbol{u}_{p,q}^{\lambda}) \right\|^2 + \lambda \|\boldsymbol{u} - \boldsymbol{u}_{p,q}^{\lambda}\|^2$$

5) $f_{\nu_p,\nu_q}^{\lambda} = \frac{\boldsymbol{u}_{p,q}^{\lambda}}{\mathcal{S}_{\mathscr{H},\lambda}(\nu_p,\nu_q)}$ is the optimal witness function of (13).

We see in that case that the optimal witness function of $\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q)$ satisfies the following identity:

$$\boldsymbol{u}_{p,q}^{\lambda} = (D(\nu_q) + \lambda I)^{-1}(\mu(\nu_p) - \mu(\nu_q)),$$

and hence regularization amounts to regularizing the KDGE. Moreover $\mathcal{S}_{\mathscr{H},\lambda}^2(\nu_p,\nu_q)$ has the interpretation of a regularized kinetic energy. We shall study the propreties of $\nabla_x u_{p,q}^{\lambda}$ as a transport map in the following Section.

5.1 Regularized Transport in RKHS: Impact of regularization on the principal Transport directions

Similarly to the un-regularized case consider $(\lambda_j, \boldsymbol{\psi}_j)$ eigenfunctions in $\mathscr{H}$ of $D(\nu_q)$ we have in this case:

$$\nabla_x u_{p,q}^{\lambda}(x) = \sum_{j=1}^m \frac{1}{\lambda_j + \lambda} \langle \boldsymbol{\psi}_j, \mu(\nu_p) - \mu(\nu_q)\rangle \nabla_x \psi_j(x),$$

hence we see that regularization is spectral filtering the principal transport directions $\nabla_x \psi_j(x)$ weighing down small eigenvalues. Hence the impact of regularization here is similar to spectral filtering principal directions of the covariance matrix in kernel PCA, but here it is filtering principal transport directions $\nabla_x \psi_j$.

6 Empirical Regularized Kernel Sobolev Discrepancy and Generalization Bounds

We define below the Empirical Regularized Kernel Sobolev Discrepancy $\hat{\mathcal{S}}_{\mathscr{H},\lambda}(\hat{\nu}_p, \hat{\nu}_q)$ for empirical measures $\hat{\nu}_p, \hat{\nu}_q$. We then give generalization bounds, i.e finite sample bounds on its convergence

convergence to the Expected Kernelized Sobolev Discrepancy $\mathcal{S}_{\mathcal{H}}(\nu_p, \nu_q)$ and the Sobolev Discrepancy $\mathcal{S}(\nu_p, \nu_q)$. We then give a closed form solution of the empirical critic of the Kernelized Sobolev Discrepancy.

Definition 5 (Regularized Empirical Kernelized Sobolev Discrepancy). *Let $\{x_i, i = 1 \ldots N, x_i \sim \nu_p\}$, and $\{y_j, j = 1 \ldots M, y_j \sim \nu_q\}$, be samples from ν_p and ν_q respectively. Let $\hat{\nu}_p(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$ and $\hat{\nu}_q(y) = \frac{1}{M} \sum_{i=1}^{M} \delta(y - y_i)$. We define the regularized empirical Kernelized Sobolev Discrepency as follows:*

$$\hat{\mathcal{S}}_{\mathcal{H},\lambda}(\hat{\nu}_p, \hat{\nu}_q) = \sup_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} f(x_i) - \frac{1}{M} \sum_{i=1}^{M} f(y_j) : \frac{1}{M} \sum_{j=1}^{M} \|\nabla_x f(y_j)\|^2 + \lambda \|f\|_{\mathcal{H}}^2 \leq 1 \right\} \quad (15)$$

achieved at $f_{\hat{\nu}_p, \hat{\nu}_q}^{\lambda}$.

Similarly to the expected case the following lemma characterizes the witness function of $\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q)$:

Lemma 3. *1)We have the following unconstrained equivalent form:*

$$\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q) = \sup_{u \in \mathcal{H}} \left\{ \frac{2}{N} \sum_{i=1}^{N} u(x_i) - \frac{2}{M} \sum_{i=1}^{M} u(y_j) - \frac{1}{M} \sum_{j=1}^{M} \|\nabla_x u(y_j)\|^2 - \lambda \|u\|_{\mathcal{H}}^2 \leq 1 \right\} \quad (16)$$

2) The optimal witness function of $\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q)$ satisfies: $\hat{u}_{p,q}^{\lambda} = (\hat{D}(\nu_q) + \lambda I)^{-1}(\hat{\mu}(\nu_p) - \hat{\mu}(\nu_q))$, where $\hat{D}(\nu_q) = \frac{1}{M} \sum_{j=1}^{M} [J\Phi(y_j)]^{\top} J\Phi(y_j)$ is the empirical KDGE and $\hat{\mu}(\nu_p) = \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i)$ and $\hat{\mu}(\nu_q) = \frac{1}{M} \sum_{j=1}^{M} \Phi(y_j)$ the empirical KMEs.

3) $\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q) = \left\| (\hat{D}(\nu_q) + \lambda I)^{-\frac{1}{2}}(\hat{\mu}(\nu_p) - \hat{\mu}(\nu_q)) \right\|^2.$

Proof. Apply proposition 1 for $\hat{\nu}_p(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$ and $\hat{\nu}_q(y) = \frac{1}{M} \sum_{i=1}^{M} \delta(y - y_i)$ □

6.1 Convergence analysis

In this Section we want first a comparison inequality between the squared Kernel Sobolev Discrepancy $\mathcal{S}_{\mathcal{H}}^2(\nu_p, \nu_q)$ and the squared empirical Regularized Sobolev Discrepancy $\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\nu_p, \nu_q)$, through their respective witness functions $u_{p,q}^{\mathcal{H}}$ and $\hat{u}_{p,q}^{\lambda}$. The following Lemma establishes this relation:

Lemma 4 (comparison inequalities). *Assume $\mathcal{S}^2(\nu_p, \nu_q) < \infty$. 1) Approximation error of $W_0^{1,2}(\mathcal{X}, \nu_q)$ in $\mathcal{H}$:*

$$|\mathcal{S}^2(\nu_p, \nu_q) - \mathcal{S}_{\mathcal{H}}^2(\nu_p, \nu_q)| \leq \inf_{u \in \mathcal{H}} \int_{\mathcal{X}} \|\nabla_x u(x) - \nabla_x u_{p,q}(x)\|^2 q(x) dx$$

2) Statistical Error, approximation with samples. Note $\delta = \mu(\nu_p) - \mu(\nu_q)$ and $\hat{\delta} = \hat{\mu}(\nu_p) - \hat{\mu}(\nu_q)$. We have:

$$|\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q) - \mathcal{S}_{\mathcal{H}}^2(\nu_p, \nu_q)| \leq \left\| \delta - \hat{\delta} \right\| \left\| \hat{u}_{p,q}^{\lambda} \right\| + (1 + \lambda) \left\| \hat{u}_{p,q}^{\lambda} \right\|^2 \left\| D(\nu_q) - \hat{D}(\nu_q) \right\|_{op}$$

$$+ \left\| \sqrt{D(\nu(q))}(\hat{u}_{p,q}^{\lambda} - u_{p,q}^{\mathcal{H}}) \right\|^2$$

Assume $M = N$. Using classical concentration results for example Theorem 4 in [7] one can show that :

$$|\hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q) - \mathcal{S}_{\mathcal{H}}^2(\nu_p, \nu_q)| \leq C \left(\frac{1}{N} + \mathscr{A}(\lambda) \right),$$

with $\lim_{\lambda \to 0} \mathscr{A}(\lambda) = 0$, and hence $|\mathcal{S}^2(\nu_p, \nu_q) - \hat{\mathcal{S}}_{\mathcal{H},\lambda}^2(\hat{\nu}_p, \hat{\nu}_q)| \leq$ $\underbrace{C \left(\frac{1}{N} + \mathscr{A}(\lambda) \right)}_{\text{Statistical Error}} + \underbrace{\inf_{u \in \mathcal{H}} \int_{\mathcal{X}} \|\nabla_x u(x) - \nabla_x u_{p,q}(x)\|^2 q(x) dx}_{\text{Approximation error}}.$ The error is therefore domi-

nated by the approximation error and the expressive power of the finite dimensional RKHS.

7

References

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

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

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

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

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

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

[7] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 2007.

A Proofs: Sobolev Discrepancy

Proof of Proposition 1 . 1) Consider the solution $u_{p,q}$ of the following PDE:

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

physically this PDE means that we are moving the mass q to p following the flows of a velocity field given by $\nabla_x u$. For any differentiable function f that vanishes on the boundary of $\mathcal{X}$, with $\|f\|_{\dot{H}(\nu_q)} \leq 1$, we have by integrating (4):

$$
\begin{aligned}
\mathbb{E}_{x\sim\nu_p} f(x) - \mathbb{E}_{x\sim\nu_q} f(x) &= \mathbb{E}_{x\sim\nu_q} \langle \nabla_x f(x), \nabla_x u_{p,q}(x)\rangle \\
&\leq \|\nabla_x h\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)} \|\nabla_x u_{p,q}\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)}
\end{aligned}
$$

Let $f^*_{\nu_p,\nu_q} = \dfrac{u_{p,q}}{\|\nabla_x u_{p,q}\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)}}$, we have:

$$\mathbb{E}_{x\sim\nu_p} f^*_{\nu_p,\nu_q}(x) - \mathbb{E}_{x\sim\nu_q} f^*_{\nu_p,\nu_q}(x) = \|\nabla_x u_{p,q}\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)} = \sqrt{\int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x)dx,}$$

It follows that for any feasible f of Problem (1) we have:

$$\mathbb{E}_{x\sim\nu_p} f(x) - \mathbb{E}_{x\sim\nu_q} f(x) \leq \|\nabla_x u_{p,q}\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)} = \mathbb{E}_{x\sim\nu_p} f^*_{\nu_p,\nu_q}(x) - \mathbb{E}_{x\sim\nu_q} f^*_{\nu_p,\nu_q}(x) = \mathcal{S}(\nu_p,\nu_q)$$

It follows therefore that the optimal witness function of (1) is $f^*_{\nu_p,\nu_q} = \dfrac{u_{p,q}}{\|\nabla_x up,q\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)}}$, where $u_{p,q}$ is solution of the advection PDE given in (4). The optimal value value of (1) is:

$$S(\nu_p,\nu_q) = \sqrt{\int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x)dx.}$$

2) Let $v : \mathcal{X} \to \mathbb{R}^d$ we claim:

$$\mathcal{S}(\nu_p,\nu_q) = \inf_{v} \left\{ \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} \text{ subject to}: p(x) - q(x) = -\text{div}(q(x)v(x)) \right\} (P) \quad (17)$$

Set the Lagrangian, and assume $f|_{\partial\mathcal{X}} = 0$

$$
\begin{aligned}
\mathcal{L}(v,f) &= \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} + \int_{\mathcal{X}} f(x)(p(x) - q(x) + div(qv(x)))dx \\
&= \int_{\mathcal{X}} f(x)(p(x) - q(x))dx + \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} - \int_{\mathcal{X}} \langle v(x), \nabla_x f(x)\rangle q(x)dx
\end{aligned}
$$

By the convexity of the problem we have that the primal formulation (P) is equal to:

$$\sup_u \inf_v \mathcal{L}(v,f) = \sup_u \left\{ \int_{\mathcal{X}} f(x)(p(x)-q(x))dx + \inf_v \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} - \int_{\mathcal{X}} \langle v(x), \nabla_x f(x)\rangle q(x)dx \right\}$$

Note that we have two cases:

$$\inf_v \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} - \int_{\mathcal{X}} \langle v(x), \nabla_x f(x)\rangle q(x)dx = -\infty \text{ if } \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 q(x)dx} > 1$$

and

$$\inf_v \sqrt{\int_{\mathcal{X}} \|v(x)\|^2 q(x)dx} - \int_{\mathcal{X}} \langle v(x), \nabla_x f(x)\rangle q(x)dx = 0 \text{ if } \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 q(x)dx} \leq 1.$$

Hence we have the primal equal to :

$$
\begin{aligned}
(P) &= \sup_{f, \sqrt{\int_{\mathcal{X}} \|\nabla_x f(x)\|^2 q(x)dx} \leq 1, \ f|_{\partial\mathcal{X}}=0} \int_{\mathcal{X}} f(x)(p(x) - q(x))dx \\
&= S(\nu_p,\nu_q) \\
&= (D).
\end{aligned}
$$

Let u the solution of (4) it follows that the solution $v^* = \nabla_x u_{p,q}$, first it is feasible and
$$\sqrt{\int_{\mathcal{X}} \|v^*(x)\|^2 q(x)dx} = \|\nabla_x u_{p,q}\|_{\mathscr{L}_2^{\otimes d}(\mathcal{X},\nu_q)} = \mathcal{S}(\nu_p, \nu_q).$$

$\square$

Proof of Proposition 2. For the upper bound $W_2(\nu_q, \nu_p) \le 2\mathcal{S}(\nu_p, \nu_q)$ the proof is given in Lemma 1 of [4] :
$$\mathscr{W}_2(\nu_q, \nu_p) \le 2\|\nu_p - \nu_q\|_{\dot{H}^{-1}(\nu_q)} = 2\mathcal{S}(\nu_p, \nu_q).$$

For the lower bound we adapt the proof of [5] given for $\|.\|_{\dot{H}^{-1}(\mathcal{X})}$. Let $f^*_{\nu_p,\nu_q}$ be the optimal Sobolev witness function we have $\left\|\nabla_x f^*_{\nu_p,\nu_q}\right\|_{\mathscr{L}_2(\mathcal{X},\nu_q)} = 1$, and $f^*_{\nu_p,\nu_q}|_{\partial\mathcal{X}} = 0$. Let (μ_{q_t}, v_t) the solutions of the dynamic formulation of Benamou and Bernier. Note that μ_{q_t} are absolutely continuous and their densities remain bounded from above and below by the same constants (see Prop 7.29 and Prop 7.30 Santambrogio book).

$$
\begin{aligned}
\mathcal{S}(\nu_p, \nu_q) &= \int_{\mathcal{X}} f^*_{\nu_p,\nu_q} d(\nu_p - \nu_q) = \int_0^1 \frac{d}{dt}\left(\int_{\mathcal{X}} f^*_{\nu_p,\nu_q}(x) d\mu_{q_t}(x)\right) dt \\
&= \int_0^1 \int_{\mathcal{X}} \left\langle \nabla_x f^*_{\nu_p,\nu_q}(x), v_t(x) \right\rangle q_t(x) dx dt \\
&\le \left(\int_0^1 \int_{\mathcal{X}} \left\|\nabla_x f^*_{\nu_p,\nu_q}(x)\right\|^2 q_t(x) dx dt\right)^{\frac{1}{2}} \left(\int_0^1 \int_{\mathcal{X}} \|v_t(x)\|^2 q_t(x) dx dt\right)^{\frac{1}{2}} \\
&= \left(\int_0^1 \int_{\mathcal{X}} \left\|\nabla_x f^*_{\nu_p,\nu_q}(x)\right\|^2 q(x) \frac{q_t(x)}{q(x)} dx dt\right)^{\frac{1}{2}} W_2(\nu_q, \nu_p) \\
&\le \sqrt{\frac{b}{a}} \left(\int_{\mathcal{X}} \left\|\nabla_x f^*_{\nu_p,\nu_q}(x)\right\|^2 q(x) dx\right)^{\frac{1}{2}} W_2(\nu_q, \nu_p) \\
&= \sqrt{\frac{b}{a}} W_2(\nu_q, \nu_p).
\end{aligned}
$$

$\square$

Proof of Lemma 1. Let $u_{p,q}$ be the solution of the following PDE:
$$p(x) - q(x) = -div(q(x)\nabla_x u_{p,q}(x)), \tag{18}$$
with boundary condition $\langle \nabla_x u_{p,q}(x), n(x)\rangle = 0$ on $\partial\mathcal{X}$. We know that :
$$\mathcal{S}^2(\nu_p, \nu_q) = \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx.$$

Step 1. Setting $u = u_{p,q}$. Let us first show that
$$L(u_{p,q}) = 2\int_{\mathcal{X}} u_{p,q}(x)(p(x) - q(x))dx - \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx = \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx.$$

$$
\begin{aligned}
L(u_{p,q}) &= -2\int_{\mathcal{X}} u_{p,q}(x) div(q(x)\nabla_x u_{p,q}(x)) - \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx \text{ (Using Equation (18))} \\
&= -2\left(\int_{\mathcal{X}} -\|\nabla_x u_{p,q}(x)\|^2 q(x) dx\right) - \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx \text{ (Divergence Theorem)} \\
&= \int_{\mathcal{X}} \|\nabla_x u_{p,q}(x)\|^2 q(x) dx \\
&= \mathcal{S}^2(\nu_p, \nu_q).
\end{aligned}
$$

Step 2. Let us show that $u_{p,q}$ is an optimizer of this loss. Meaning that for all u we have $L(u) \leq L(u_{p,q})$.

$$
\begin{aligned}
L(u) &= 2 \int_{\mathcal{X}} u(x)(p(x) - q(x))dx - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x)dx \\
&= -2 \int_{\mathcal{X}} u(x)div(q(x)\nabla_x u_{p,q}) - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x)dx \text{ (Using Equation (18))} \\
&= 2 \int_{\mathcal{X}} \langle \nabla_x u(x), u_{p,q}(x) \rangle q(x)dx - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x)dx \text{ (Divergence Theorem)} \\
&= 2 \int_{\mathcal{X}} \langle \nabla_x u(x), u_{p,q}(x) \rangle q(x)dx - \int_{\mathcal{X}} \|\nabla_x u(x)\|^2 q(x)dx - L(u_{p,q}) + L(u_{p,q}) \\
&= L(u_{p,q}) + \int_{\mathcal{X}} \left(2 \langle \nabla_x u(x), u_{p,q}(x) \rangle - \|\nabla_x u(x)\|^2 - \|\nabla_x u_{p,q}(x)\|^2 \right) q(x)dx \\
&= L(u_{p,q}) - \int_{\mathcal{X}} \|\nabla_x u(x) - \nabla_x u_{p,q}(x)\|^2 q(x)dx \\
&\leq L(u_{p,q}) = \mathcal{S}^2(\nu_p, \nu_q).
\end{aligned}
$$

with equality when $u = u_{p,q}$. $\qquad\square$

B Proofs: Kernel Sobolev Discrepancy

Proof of Proposition 3. It is easy to see that for $f \in \mathcal{H}$,
$$
\mathbb{E}_{x \sim \nu_p} f(x) - \mathbb{E}_{x \sim \nu_q} f(x) = \langle \boldsymbol{f}, \mu(\nu_p) - \mu(\nu_q) \rangle,
$$
where $\mu(\nu_p)$ and $\mu(\nu_q)$ are the KME of ν_p and ν_q. On the other hand:
$$
\mathbb{E}_{x \sim \nu_q} \|\nabla_x f(x)\|^2 = \langle \boldsymbol{f}, D(\nu_q)\boldsymbol{f} \rangle
$$
where $D(\nu_q)$ is the KDGE of ν_q (as defined in Equation (10)). Hence we have under the assumption that $D(\nu_q)$ is invertible:

$$
\begin{aligned}
\mathcal{S}_{\mathcal{H}}(\nu_p, \nu_q) &= \sup_{f \in \mathcal{H}, \langle f, D(\nu_q)f \rangle_{\mathcal{H}} \leq 1} \langle f, \mu(\nu_p) - \mu(\nu_q) \rangle \\
&= \sup_{\boldsymbol{g}\|\boldsymbol{g}\| \leq 1} \left\langle \boldsymbol{g}, D^{-\frac{1}{2}}(\nu_q)\left(\mu(\nu_p) - \mu(\nu_q)\right) \right\rangle \\
&= \left\| D^{-\frac{1}{2}}(\nu_q)\left(\mu(\nu_p) - \mu(\nu_q)\right) \right\|,
\end{aligned}
$$

and
$$
\boldsymbol{f}_{\nu_p,\nu_q}^{\mathcal{H}} = \frac{1}{\mathcal{S}_{\mathcal{H}}(\nu_p, \nu_q)} [D(\nu_q)]^{-1} \left(\mu(\nu_p) - \mu(\nu_q)\right),
$$
and $\boldsymbol{u}_{p,q}^{\mathcal{H}} = [D(\nu_q)]^{-1} \left(\mu(\nu_p) - \mu(\nu_q)\right).$

$\qquad\square$

Proof of Proposition 4. We have: $\mathcal{S}_{\mathcal{H}}(\nu_p, \nu_q) \leq \mathcal{S}(\nu_p, \nu_q)$ $\qquad\square$

C Proofs:Regularized Kernel Sobolev Discrepancy

Proof of Theorem 1. Let $\boldsymbol{u}_{p,q}^{\lambda} \in \mathcal{H}$ be the solution of :
$$
(D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^{\lambda} = \mu(\nu_p) - \mu(\nu_q)
$$
We know that the solution of (13) satisfies (for a proof it is similar to 3 just adding the regularization $\lambda > 0$):

$$
\begin{aligned}
\mathcal{S}_{\mathcal{H},\lambda}^2(\nu_p, \nu_q) &= \langle \mu(\nu_p) - \mu(\nu_q), (D(\nu_q) + \lambda I)^{-1}(\mu(\nu_p) - \mu(\nu_q)) \rangle_{\mathcal{H}} \\
&= \langle \boldsymbol{u}_{p,q}^{\lambda}, D(\nu_q)\boldsymbol{u}_{p,q}^{\lambda} \rangle + \lambda \|\boldsymbol{u}_{p,q}^{\lambda}\|^2 \\
&= \int_{\mathcal{X}} \|\nabla_x u_{p,q}^{\lambda}(x)\|^2 q(x)dx + \lambda \|\boldsymbol{u}_{p,q}^{\lambda}\|^2.
\end{aligned}
$$

Let

$$L(u, \lambda) = 2 \langle \boldsymbol{u}, \mu(\nu_p) - \mu(\nu_q) \rangle - \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u} \rangle$$

Step 1.

$$\begin{aligned}
L(u_{p,q}^\lambda, \lambda) &= 2 \langle \boldsymbol{u}_{p,q}^\lambda, \mu(\nu_p) - \mu(\nu_q) \rangle - \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle \\
&= 2 \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle - \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle \\
&= \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle \\
&= \mathcal{S}_{\mathscr{H},\lambda}^2 (\nu_p, \nu_q).
\end{aligned}$$

Step 2. Let us show that $\boldsymbol{u}_{p,q}^\lambda$ is an optimizer of this loss. Meaning that for all u we have $L(u, \lambda) \leq L(u_{p,q}^\lambda, \lambda)$.

$$\begin{aligned}
L(u, \lambda) - L(u_{p,q}^\lambda, \lambda) &= 2 \langle \boldsymbol{u}, \mu(\nu_p) - \mu(\nu_q) \rangle - \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u} \rangle - \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle \\
&= 2 \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle - \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u} \rangle - \langle \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)\boldsymbol{u}_{p,q}^\lambda \rangle \\
&= - \langle \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda, (D(\nu_q) + \lambda I)(\boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda) \rangle \\
&= - \langle \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda, D(\nu_q)(\boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda) \rangle - \lambda \left\| \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda \right\|^2 \\
&= - \int_{\mathcal{X}} \left\| \nabla_x u(x) - \nabla_x u_{p,q}^\lambda(x) \right\|^2 q(x) dx - \lambda \left\| \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda \right\|^2 \\
&= - \left\| (D(\nu_q))^{\frac{1}{2}} (\boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda) \right\|^2 - \lambda \left\| \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda \right\|^2 \\
&\leq 0.
\end{aligned}$$

with equality when $u = \boldsymbol{u}_{p,q}^\lambda$. 1) to 5) follow immediately from the proof above. $\qquad\square$

D Convergence

Proof of Lemma 4. 1) We have from Lemma 1 (Equation (8)):

$$\begin{aligned}
0 \leq \mathcal{S}^2(\nu_p, \nu_q) - \mathcal{S}_{\mathscr{H}}^2(\nu_p, \nu_q) &= \int_{\mathcal{X}} \left\| \nabla_x u_{p,q}^{\mathscr{H}}(x) - \nabla_x u_{p,q}(x) \right\|^2 q(x) dx \\
&\leq \inf_{u \in \mathscr{H}} \int_{\mathcal{X}} \left\| \nabla_x u(x) - \nabla_x u_{p,q}(x) \right\|^2 q(x) dx
\end{aligned}$$

2) Let

$$L(u, \lambda) = 2 \langle \boldsymbol{u}, \mu(\nu_p) - \mu(\nu_q) \rangle - \langle \boldsymbol{u}, (D(\nu_q) + \lambda I)\boldsymbol{u} \rangle$$

and

$$\hat{L}(u, \lambda) = 2 \langle \boldsymbol{u}, \hat{\mu}(\nu_p) - \hat{\mu}(\nu_q) \rangle - \left\langle \boldsymbol{u}, (\hat{D}(\nu_q) + \lambda I)\boldsymbol{u} \right\rangle.$$

Note $\delta = \mu(\nu_p) - \mu(\nu_q)$ and $\hat{\delta} = \hat{\mu}(\nu_p) - \hat{\mu}(\nu_q)$.
From Theorem 1 point 4) we have: For any $\lambda > 0$, and any $u \in \mathscr{H}$:

$$L(u_{p,q}^\lambda, \lambda) - L(u, \lambda) = \left\| (D(\nu_q))^{\frac{1}{2}} (\boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda) \right\|^2 + \lambda \left\| \boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda \right\|^2$$

In particular for the unregularized case $\lambda_0 = 0$ we have $L(u_{p,q}^{\lambda_0}, \lambda_0) = L(u_{p,q}^{\mathscr{H}}, 0) = \mathcal{S}_{\mathscr{H}}^2(\nu_p, \nu_q)$. Hence for this particular case we have for any $u \in \mathscr{H}$

$$\mathcal{S}_{\mathscr{H}}^2(\nu_p, \nu_q) - L(u, 0) = L(u_{p,q}^{\mathscr{H}}, 0) - L(u, 0) = \left\| (D(\nu_q))^{\frac{1}{2}} (\boldsymbol{u} - \boldsymbol{u}_{p,q}^\lambda) \right\|^2$$

$$\hat{\mathcal{S}}^2_{\mathscr{H},\lambda}(\hat{\nu}_p, \hat{\nu}_q) - \mathcal{S}^2_{\mathscr{H}}(\nu_p, \nu_q) = \hat{L}(\hat{u}^\lambda_{p,q}, \lambda) - L(\hat{u}^\lambda_{p,q}, \lambda) + L(\hat{u}^\lambda_{p,q}, \lambda) - L(u^{\mathscr{H}}_{p,q}, 0)$$

$$= 2 \left\langle \hat{\boldsymbol{u}}^\lambda_{p,q}, \hat{\delta} - \delta \right\rangle - \left\langle \hat{\boldsymbol{u}}^\lambda_{p,q}, (\hat{D}(\nu_q) - D(\nu_q))\hat{\boldsymbol{u}}^\lambda_{p,q} \right\rangle$$

$$+ L(\hat{u}^\lambda_{p,q}, 0) + \lambda \left\| \hat{\boldsymbol{u}}^\lambda_{p,q} \right\|^2 - L(u^{\mathscr{H}}_{p,q}, 0)$$

$$= 2 \left\langle \hat{\boldsymbol{u}}^\lambda_{p,q}, \hat{\delta} - \delta \right\rangle - \left\langle \hat{\boldsymbol{u}}^\lambda_{p,q}, (\hat{D}(\nu_q) - D(\nu_q))\hat{\boldsymbol{u}}^\lambda_{p,q} \right\rangle + \lambda \left\| \hat{u}^\lambda_{p,q} \right\|^2$$

$$- \left\| \sqrt{D(\nu(q))}(\hat{\boldsymbol{u}}^\lambda_{p,q} - \boldsymbol{u}^{\mathscr{H}}_{p,q}) \right\|^2$$

$$|\hat{\mathcal{S}}^2_{\mathscr{H},\lambda}(\hat{\nu}_p, \hat{\nu}_q) - \mathcal{S}^2_{\mathscr{H}}(\nu_p, \nu_q)| \le \left\| \delta - \hat{\delta} \right\| \left\| \hat{\boldsymbol{u}}^\lambda_{p,q} \right\| + (1 + \lambda) \left\| \hat{\boldsymbol{u}}^\lambda_{p,q} \right\|^2 \left\| D(\nu_q) - \hat{D}(\nu_q) \right\|_{op}$$

$$+ \left\| \sqrt{D(\nu(q))}(\hat{\boldsymbol{u}}^\lambda_{p,q} - \boldsymbol{u}^{\mathscr{H}}_{p,q}) \right\|^2$$

$\square$

$$\hat{\mathcal{S}}^2_{\mathscr{H},\lambda}(\hat{\nu}_p, \hat{\nu}_q) - \mathcal{S}^2_{\mathscr{H}}(\nu_p, \nu_q) = \hat{L}(\hat{u}^\lambda_{p,q}, \lambda) - L(\hat{u}^\lambda_{p,q}, \lambda) + L(\hat{u}^\lambda_{p,q}, \lambda) - L(u^{\mathscr{H}}_{p,q}, 0)$$