X-Ray Sobolev Variational Auto-Encoders

Gabriel Turinici

X-Ray Sobolev Variational Auto-Encoders

Gabriel Turinici
Universit Paris Dauphine - PSL Research University
CEREMADE,
Place du Marechal de Lattre de Tassigny, Paris 75016, FRANCE
Gabriel.Turinici@dauphine.fr

February 1st, 2020

Abstract

The quality of the generative models (Generative adversarial networks, Variational Auto-Encoders, ...) depends heavily on the choice of a good probability distance. However some popular metrics lack convenient properties such as (geodesic) convexity, fast evaluation and so on. To address these shortcomings, we introduce a class of distances that have built-in convexity. We investigate the relationship with some known paradigms (sliced distances, reproducing kernel Hilbert spaces, energy distances). The distances are shown to posses fast implementations and are included in an adapted Variational Auto-Encoder termed X-ray Sobolev Variational Auto-Encoder (XS-VAE) which produces good quality results on standard generative datasets.

1 Introduction

Deep neural networks used as generative models are of high interest in a large array of application domains [1, 2, 3, 4]. However they come at the price of a more intricate architecture and convergence patterns than supervised networks.

The goal of generative models is to design a procedure to sample from a target probability law $\mathbb{P}^{real}$ using a dataset of available samples Y_1, ..., $Y_L \sim \mathbb{P}^{real}$ (the number L of samples is large but fixed).

One of the most used and efficient architectures are the Generative Adversarial Networks (GANs); GANs come in the form of a dual net: a generator and a discriminator, whose joint convergence was shown to pose problems, addressed in late variants (see WGAN, SWGAN, etc.). As generators deal with probability

laws, the choice of the distance used to quantify the closeness of a candidate turns out to be of critical importance. This is even more visible for Variational Auto-Encoders (VAE) that use the distance directly (i.e., not in the dual form as most of the GANs do).

A VAE has two stages: an encoder stage $E_{\theta_e}(\cdot)$ indexed by the parameters θ_e of the encoding network and a decoding network D_{θ_d}. The networks are fitted in order to satisfy two goals: first the reconstruction error $D_{\theta_d}(E_{\theta_e}(Y_k)) - Y_k$ is to be minimized over all available samples Y_k; this is usually implemented minimizing the mean square error. On the other hand, the second requirement is to minimize the mismatch between the encoded empirical distribution $E_{\theta_e}\#\mathbb{P}^{real}$ and a predefined, fixed, law $\mathbb{L}_z$ (here the symbol "$\#$" means the image of the distribution on the right with respect to the mapping on the left, also called the push-forward map).

In the generation phase one takes as input samples z from the law $\mathbb{L}_z$ and maps them through the decoding function D_{θ_d} in order to generate new samples not seen in the dataset $Y_1, ..., Y_L$.

Let Ω be the set of all possible values of the latent sample z and $\mathcal{P}(\Omega)$ the set of all probability laws on Ω; then, crucial to the VAE is the distance $d(\cdot, \cdot)$ acting on (possibly a subset of the) $\mathcal{P}(\Omega)^2$ that measures the mismatch between $E_{\theta_e}\#\mathbb{P}^{real}$ and the latent distribution $\mathbb{L}_z$. The computation of the distance d is usually difficult and in many cases one resorts to approximations.

The goal of this paper is to present a class of distances relevant to practice, easy to compute and that have built-in convexity (thus ease the minimization procedure).

Our proposal builds on several ideas: on one hand the "kernel trick" that enables to compute functional quantities on Euclidian spaces; on the other hand the "sliced distances" that average, in a sense to be made precise, the distances of projections to one-dimensional subspaces. We use the "X-ray" term instead of "slice" to eliminate the ambiguity between the Radon and X-ray transforms, the first projecting on hyper-planes (dimension $N - 1$) and the other on one-dimensional spaces (dimension 1), see [5], [6, Chapter 2] (the two coincide only in dimension $N = 2$, which is much lower than the dimension of applications we have in mind). The third and last ingredient is a Hilbert space of regular functions. The regularity has often been invoked in relation with GANs for instance in the Wasserstein GAN form (that use Lipschitz functions). We encode the smoothness of a function through a parameter $s \geq 0$ of some Sobolev space H^s of $\dot{H}^s$ (see definition in efsec:sobolev), larger s meaning smoother.

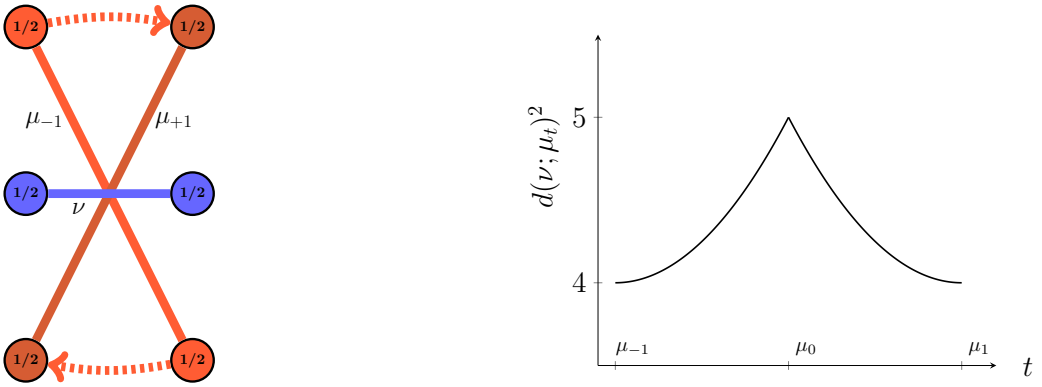

Figure 1: **Left:** Illustration of the sum of Diracs forming a geodesic (in red shades) $(\mu_t)_{t\in[-1,1]}$ such that $t \mapsto d_{W_1}(\mu_t, \nu)^2$ is severely non-convex, where ν the measure depicted in blue. **Right:** The distance squared from points μ_t on a geodesic to the target ν. The function is not convex and in particular has two local minima.

2 Desirable properties of a distance

We discuss in this section some specific properties that a distance should have in order to be suitable for use in generative models.

2.1 Convexity

A property easy to understand is the convexity; it is not a hard requirement but it certainly helps the convergence and robustness with respect to perturbations.

Let us be more clear what convexity means in a space of probability laws. Suppose a metric d is given on $\mathcal{P}(\Omega)$; then given two distributions μ_{-1} and μ_1 one can consider the geodesic μ_t starting at μ_{-1} and ending in μ_1. In an Euclidian space this would be just a straight line but in general this is not the case [6]. Take for instance the Wasserstein metric d_W used for instance in WGANs (see [4] for details concerning the definition of such metrics), and the following example (adapted from [6, page 275]): consider the geodesic $\mu_t = \frac{1}{2}\delta_{(t,2)} + \frac{1}{2}\delta_{(-t,-2)}$ linking μ_{-1} and μ_1 and the "target" law $\nu = \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(-1,0)}$. Then the distance squared from ν to the points μ_t on the geodesic is given by: $d(\nu, \mu_t)^2 = 4 + \min((1-t)^2, (1+t)^2)$, which is not convex, see Figure 1. Or, the distance squared function is an important ingredient of the many loss functions and the additional non-convexity added because of using the Wasserstein distance will surcharge even more the optimization effort.

2.2 Nonlocal calculability

Another suitable property of a distance is related to its calculability. It is very convenient to deal with distances that can be computed with as less information as possible concerning the local (thus specific and changing) properties of the distributions in the argument. For instance, in a metric space, the possibility to compute the distance to points on a geodesic with the sole information of the distance to its extremities is very practical and has been used in many works [7]. We will consider in particular the following property:

$$\begin{aligned} &\text{For any geodesic } \mu_t : [0,1] \to \mathcal{P}(\Omega) \text{ and any } \nu \in \mathcal{P}(\Omega) : \\ &\quad d^2(\nu,\mu_t) = (1-t)\cdot d^2(\nu,\mu_0) + t\cdot d^2(\nu,\mu_1) \\ &\quad -t(1-t)\cdot d^2(\mu_0,\mu_1), \forall t \in [0,1]. \end{aligned} \tag{1}$$

Although at first cryptic, this is nothing more than, for instance, the parallelogram identity in a Hilbert space $\|x+y\|^2 = 2\cdot\|x\|^2 + 2\cdot\|y\|^2 - \|x-y\|^2$ expressed at $\nu = 0$, $\mu_0 = x$, $\mu_1 = y$, $t = 1/2$. In fact (1) is satisfied in any Hilbert space because the geodesics are straight lines. For instance the space L^2 of square integrable functions is a good example, while the space L^4 of functions with finite norm $\|f\|_{L^4} = \left(\int f^4\right)^{1/4}$ does not fulfill the condition. Reciprocally, if a metric space is endowed with an algebraic operation compatible with the distance and the above identity, it can then be isometrically embedded into a Hilbert space. In practice this allows to have the following:

Proposition 1 *Let Ω a subdomain of $\mathbb{R}^N$ and (X,d) a metric space containing all Dirac masses δ_x, for all $x \in \Omega$. Then if (X,d) satisfies property* (1) *on any straight line $\mu_t = (1-t)\mu_0 + t\mu_1$ then for any $x_1, ..., x_K$, $y_1, ..., y_J \in \Omega$:*

$$\begin{aligned} &d\left(\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}, \frac{1}{J}\sum_{j=1}^{J}\delta_{y_j}\right)^2 = \frac{1}{K\cdot J}\sum_{k=1}^{K}\sum_{j=1}^{J} d(\delta_{x_k},\delta_{y_j})^2 \\ &-\frac{1}{2K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K} d(\delta_{x_k},\delta_{x_{k'}})^2 - \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J} d(\delta_{y_j},\delta_{y_{j'}})^2 \end{aligned} \tag{2}$$

Proof: follows directly from equation (1).

Proposition 1 is a powerful tool because of two reasons: first it allows to compute the distance in an easy way, second this computation is in the form of expectations (replacing the sums by expectation over a discrete uniform law; see equation (5) below for an example).

3 Relationship with the literature

This work owes much to previous advances from the literature.

Of course, first of all is the idea that a (W) GAN can be seen as minimizing some probability metric [4]. They use a Wasserstein metric which is not finally so much different from ours (see equation (12) and section 5.2.2) but whose computation requires iterations and the enforcing of some special (Lipschitz type) constraints.

Starting from this difficulty, several solutions were proposed : first the "sliced" distances, most used being the "sliced Wasserstein distance" appearing in [8, 9, 10, 11, 12] to cite but a few. Other forms of sliced distances have also been proposed, in particular in [13] authors implement a VAE using a kernelized distance; note however that in full rigor the metric they used is not a distance because of the smoothing parameter depending on the number of Dirac masses present in the distributions. However it is a good metric that allows to compute analytically the distance from a Dirac mass to a standard multi-variate normal (distance used to converge to the latent distribution).

On the other hand there is the remarkable work on the energy distance by Szekely and al. (see [14] and related references) that establish an encouraging framework; at the contrary of the first group cited, here the approach is global; note that in [15] the authors use Sobolev spaces but instead of using X-ray (or "sliced") versions they use directly the overall space $H^s(\mathbb{R}^N)$; the inconvenience of this space is that it is included in the space of continuous functions only when $s \geq N/2$ which means that, since N is large, a high regularity is required in order for the dual to contain Dirac masses. A close work is [16] which implements a particular form of Xray Sobolev distance (for $H = \dot{H}$) in the framework of a GAN; as such it requires the use of a space of features that is to be optimized.

Finally, this work can be put in the more general framework of Hilbert space embeddings (see for instance [17, Theorem 21 p.1551]) and [18] that give theoretical insights into the use of energy-type distances and relationship with MMD metrics.

3.1 Contributions of this work

This work proposes some novelties with respect to the literature that we detail below:

- first from a theoretical point of view, we introduce a novel procedure to construct a class of probability distances taking into account the regularity of the test functions; note that this regularity is precisely what delimitates for instance the Total Variation distance (continuous functions) from the Wasserstein distance (Lipschitz functions);

- in particular the "energy distance" of Szekely et al.(see [14]) is a particular member of this class; this distance has been tested extensively outside the deep learning community;

- the distances are interesting computationally because one has quasi-analytic formulas for computing the distances among sums of Dirac masses and from a sum of Dirac mass to the standard normal multi-variate distribution; furthermore some aysmptotic expansions (giving good results in practice) are proposed;

- we propose an adapted Variational Auto-Encoder (termed XS-VAE to recall that the distance used is in the proposed class) that obtains good results on standard datasets. In particular the distance is seen to be a reliable proxy for other metrics (including Sliced Wasserstein and Cramer-Wold). Note that in practice it is difficult to compare with non-sliced VAEs because the distances, exact to the extent needed to make meaningful comparisons, are difficult to obtain in high dimensional latent spaces. This comparison is for instance the approach used in other, not directly related, endeavors, see [19]).

4 Theoretical results

We refer the reader to A for the definition of Sobolev spaces $H^s(\mathbb{R})$ and $\dot{H}^s(\mathbb{R})$.

4.1 Construction of the dual X-ray distance

We make precise in this section the construction of the X-ray distance; we recall that this is another version of the "sliced" idea used in the literature [13, 9, 12, 8, 11, 10, 16] but, to eliminate any ambiguity with the Radon transform arriving in dimension $N > 2$ we keep the X-ray denomination as in the original papers [5]. We suppose thus having at our disposal a distance between two one-dimensional distributions (for instance coming from the dual of a Hilbert space $\mathcal{H}$ containing continuous functions as in Theorem 1). Note that here $\mathcal{H}$ is a generic name and it has to be instantiated (for instance $\mathcal{H} = H^1$).

Then the X-ray distance corresponding to $\mathcal{H}$ is the mean value of the distances of projected distribution over all directions on the d-dimensional unit sphere S^{N-1}; the formal definition is:

$$d_{X\mathcal{H}}(\mu,\nu)^2 = \fint_{S^{N-1}} \|\theta\#\mu - \theta\#\nu\|^2_{\mathcal{H}'} d\theta. \tag{3}$$

The norm is calculated in the dual space $\mathcal{H}'$ of $\mathcal{H}$, where the measures $\theta\#\mu$ and $\theta\#\nu$ belong. Recall that the projection $\theta\#\mu$ of the measure μ on the direction θ has, when μ is a sum of Dirac masses, the simple expression $\theta\#\left(\sum_{k=1}^K \delta_{x_k}\right) = \sum_{k=1}^K \delta_{\langle x_k,\theta\rangle}$.

When $\|\delta_x - \delta_y\|_{\mathcal{H}'}$ only depends on $|x-y|$ (as is the case for translation invariant norms and in particular for the Sobolev spaces in A) then there exists a function g such that for any $x, y \in \Omega$: $d_{X\mathcal{H}}(\delta_x, \delta_y)^2 = g(\|x-y\|)$. In this case all that is required to compute the distance $d_{X\mathcal{H}}$ is the knowledge of a real function g of one variable. For instance formula in equation (2) becomes:

$$\begin{aligned} & d_{X\mathcal{H}}\left(\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}, \frac{1}{J}\sum_{j=1}^{J}\delta_{y_j}\right)^2 \\ & = \sum_{k=1}^{K}\sum_{j=1}^{J} g(\|x_k - y_j\|) - \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K} g(\|x_k - x_{k'}\|)}{2K^2} \\ & - \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J} g(\|y_j - y_{j'}\|)}{2J^2}. \end{aligned} \tag{4}$$

4.2 Explicit formulas for the distance

With these provisions we can prove the following

Theorem 1 *Let $M > 0$ (M can be $+\infty$) and Ω the ball of radius M in $\mathbb{R}^N$ (whole $\mathbb{R}^N$ if $M = \infty$). Let $(\mathcal{H}, \|\cdot\|_H)$ be a Hilbert space of real functions of one variable (such as $H^s(\mathbb{R})$ or $\dot{H}^s(\mathbb{R})$) on the domain $]-M, M[$ such that $\mathcal{H}$ is included in the set of continuous functions $C^0(\Omega)$. Denote by $(\mathcal{H}', \|\cdot\|_{\mathcal{H}'})$ the dual of H, $\mathcal{P}(\Omega)$ the set of probability laws on Ω and $d_{X\mathcal{H}}$ the distance on $\mathcal{P}(\Omega)$ associated to $\|\cdot\|_{\mathcal{H}'}$ as described in section 4.1. Then $d_{X\mathcal{H}}$ is such that:*

1. *any line $\mu_t = (1-t)\mu_0 + t\mu_1 \subset \mathcal{P}(\Omega)$, $t \in [0,1]$ is a geodesic and $d_{X\mathcal{H}}$ is convex on μ_t;*
2. *for any $\nu \in \mathcal{P}(\Omega)$ the distance $d_{X\mathcal{H}}$ satisfies* (1)*;*
3. *the distance $d_{X\mathcal{H}}$ satisfies relation* (2)*.*

In particular, up to a multiplicative constant, for $\mathcal{H} = \dot{H}^1$:

$$\begin{aligned} & d_{X\dot{H}^1}(\mu,\nu)^2 = \mathbb{E}_{X\sim\mu, Y\sim\nu, X\perp\!\!\!\perp Y}\|X - Y\| \\ & - \frac{\mathbb{E}_{X,X'\sim\mu, X\perp\!\!\!\perp X'}\|X - X'\| + \mathbb{E}_{Y,Y'\sim\nu, Y\perp\!\!\!\perp Y'}\|Y - Y'\|}{2}, \end{aligned} \tag{5}$$

where X, X', Y, Y' are independent random variables, first two with distribution μ and the last two with distribution ν.

Proof: To prove the first two points it is enough to recall that property 1 is true in a Hilbert space because there the distance (squared) is the norm (squared) and the identity follows from the expansion using the scalar product. It remains to see that the construction in section 4.1 uses the space of square integrable functions from S^{N-1} to $\mathcal{H}'$: to each distribution μ we associate the function $f : S^{N-1} \to \mathcal{H}'$ by the relation $\theta \in S^{N-1} \mapsto \theta\#\mu \in \mathcal{H}'$. Since the dual $\mathcal{H}'$ of $\mathcal{H}$ is also a Hilbert space, the set of such functions f is a subset of $L^2(S^{N-1}, \mathcal{H}')$ which is a Hilbert space with the usual scalar product $\langle f, g\rangle = \int_{S^{N-1}} \langle f(\theta), g(\theta)\rangle_{\mathcal{H}',\mathcal{H}'} d\theta$.

The third point follows as in Proposition 1.

Remark 1 *In general, the procedure in section 4.1 allows to resume the distance to two functions g_{xy} and g_{xn} such that*

$$d_{X\mathcal{H}}(\delta_x, \delta_y)^2 = g_{xy}(\|x - y\|),\ \forall x, y \in \Omega, \tag{6}$$

$$d_{X\mathcal{H}}(\delta_x, N(0, I))^2 = g_{xn}(\|x\|),\ \forall x \in \Omega. \tag{7}$$

These functions can be used in the XS - Variational Auto-Encoder as detailed in section 5. Note that, e.g., contrary to sliced Wasserstein implementations, here there is no need to discretize the directions of the N-dimensional sphere S^{N-1}. The precise formulas and approximation of g_{xy} and g_{xn} when $\mathcal{H} = \dot{H}^1$ are the object of the B.

5 XSVAE: X-Ray Sobolev Variational Auto-Encoder

5.1 Algorithm

Enabled by the previous results, we propose a new type of VAE using the above distances. Compared with a GAN the VAE has the advantage to use a fixed reference latent distribution and does not need to look for a suitable "feature space" to express the distance in. This means that, as in [13], one can pre-compute the distance from a Dirac mass to a target latent distribution (here a standard multivariate normal) which speeds up even more the computations.

In order to compare with results from the literature we use the space $\mathcal{H} = \dot{H}$ that has been studied before and that has also interesting properties, cf. equation (12).

Let us fix the following notations: θ_e are the parameters of the encoder network and θ_d the parameters of the decoder. The encoding is a (parametrized) function transforming some real sample $X \in \mathbb{R}^I$ to the latent space $\mathbb{R}^N$, i.e. $E_{\theta_e}(X) \in \mathbb{R}^N$. The goal of this part is to have the distribution $\mathbb{P}^{real}$ transported to the latent

distribution (standard multi-variate normal) on the latent space; the corresponding part of the loss functional is $d_{X\dot{H}^1}(E_{\theta_e}\#\mathbb{P}^{real}, N(0,I))^2$. The decoding part takes $z \in \mathbb{R}^N$ and through the decoding function D_{θ_d} maps it to the initial space $\mathbb{R}^I$; the goal is to provide accurate reconstruction of the real samples and corresponds to the minimization of the mean squared error $\mathbb{E}_{X\sim\mathbb{P}^{real}}\|[D_{\theta_d}\circ E_{\theta_e}]\#(X) - X\|^2$. We obtain the Algorithm A1.

Algorithm A1 Xray-Sobolev Variational Auto-Encoder (XSVAE)

procedure XSVAE
  • Sets batch size K, latent dimension N, $\lambda \geq 0$
  • compute constants c_{N0}, c_{N1} from equation (19)
  while (stopping not reached) **do**
    • Sample $X_1, ..., X_k \sim \mathbb{P}^{real}$ (i.i.d).
    • propagate the real sample through the
    encoding network;
    • Compute the latent loss: $Loss_{lat}(\theta_e, \theta_d)$
    $= c_{N0} + \frac{\sum_{k=1}^K \sqrt{\|E_{\theta_e}(X_k)\|^2 + c_{N1}}}{K}$
    $- \frac{\sum_{k=1}^K \sum_{k'=1}^K \|E_{\theta_e}(X_k) - E_{\theta_e}(X_{k'})\|}{2K^2}$.
    • propagate the real sample through the
    decoding network;
    • Compute the reconstruction loss:
    $Loss_{rec}(\theta_e, \theta_d)$
    $= \frac{1}{K}\sum_{k=1}^K \|D_{\theta_d}(E_{\theta_e}(X_k)) - X_k\|^2$
    • Compute the global loss: $L_{global}(\theta_e, \theta_d)$
    $= Loss_{rec}(\theta_e, \theta_d) + \lambda Loss_{lat}(\theta_e, \theta_d)$
    • Backpropagate $L_{global}(\theta_e, \theta_d)$ and update
    parameters θ_e, θ_d (a minimization step).
  end while
end procedure

5.2 Numerical results

We tested the XSVAE on several datasets from the literature. The goal here is to show that this procedure is comparable with other probability distances used in generative networks while still benefiting from a fast computation and no hyper-parameters to choose (except latent space size and batch size that all VAEs deal with).

The code is available in the supplementary material and the comparison is made with the algorithm from [13]. The datasets are MNIST, Fashion-MNIST

and CIFAR10 and the encoder-decoder architecture is taken from the reference (and recalled in C).

In all cases we used the additive version of the loss functional because the "log" introduced in [13] for the distance term would prevent from having a convex loss (the "log" being concave). In practice the scaling constant λ can either be choosen by trial and error or by running first two optimizations, one with only reconstruction loss and the other with only latent loss and take the quotient of the standard deviations in the oscillations seen in the result. In practice it has been set to $\lambda = 100$ for all tests.

5.2.1 Relevance of both loss terms

We first test whether in the XS-VAE both loss terms are effective. We only show the results for MNIST dataset but the conclusions are similar for all datasets. Namely we consider three runs:
• optimization of the reconstruction loss only ($\lambda = 0$);
• optimization of the latent loss only;
• full optimization ($\lambda = 100$).

The results are in Figure 2. The first situation results in good reconstruction but poor generation while the second generates "white noise like" images. The advantage of having both terms is obvious from the figure.

When only optimizing the reconstruction loss the reconstruction error drops from 39.5 to around 11.5. When optimizing the latent part only the reconstruction loss remains at high values (depending on the run between 40 and 180) and the latent distance reaches small values (around $1.e-4$ and below). When using both terms, the reconstruction error drops down to around 11.5 and the distance around $5e-3$, depending on the run.

This behavior remains visible in all tests we performed.

5.2.2 Comparison with other "sliced" distances

One of the main interests of the distance that we introduce is to be able to provide a computational efficient alternative to previous metrics such as the sliced Wasserstein [8, 9, 10, 11, 12] (referenced as "SW" in the following) and the Cramer-Wold distances [13] (named "CW"). As such it is interesting to know whether optimizing the X-ray Sobolev distance (referenced as "XS") is a proxy to minimizing the SW or CW distances. In fact, evidence is already present in the literature that optimizing the CW distance and SW distances is similar so for clarity we will only compare with a distance at a time.

A comparaison between the XS and CW distances on the Fashion-MNIST

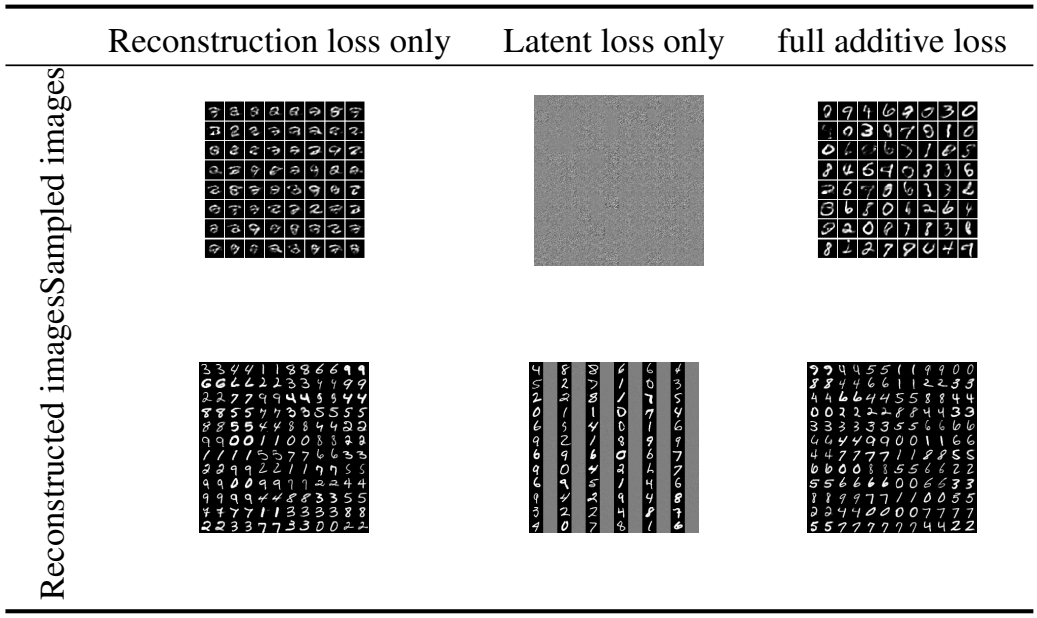

Figure 2: Results for section 5.2.1 on the MNIST dataset. Both the reconstruction cost and the latent cost are required for good generative quality.

dataset is presented in Figure 4 (see Figure 3 for illustrative results on Fashion-MNIST).

Comparaisons between the SW and XS distances on MNIST and CIFAR10 are presented in Figures 5 and 6. Both show that the XS distance is a practical proxy for both the SW and CW distances.

6 Concluding remarks

We introduce in this work a class of probability distances to be used in generative modeling; all members of the class share the important properties of convexity and fast evaluation which were not always present in previous works (Wasserstein, sliced distances, ...). Each distance corresponds to a Hilbert space of functions (Sobolev spaces in this work) and is constructed using the X-ray transform.

To illustrate the effectiveness of the metrics, we consider the particular case of the Sobolev space $\dot{H}^{-1}$ which gives a distance already present in the literature, for which we derive and use novel fast evaluation formulas.

The resulting procedure, called XS-VAE (to recall the X-ray Sobolev construction) is shown to perform well on standard datasets.

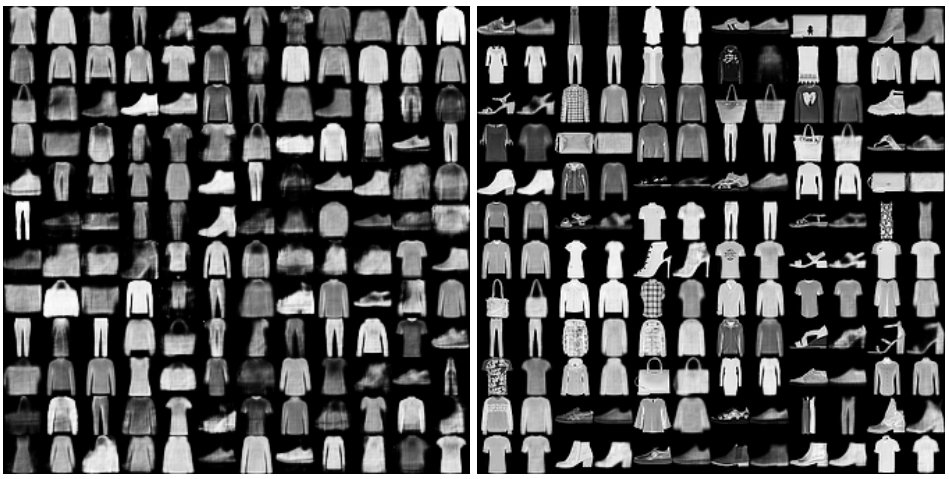

Figure 3: Fashion-MNIST dataset. **Left:** generated samples. **Right:** reconstruction quality.

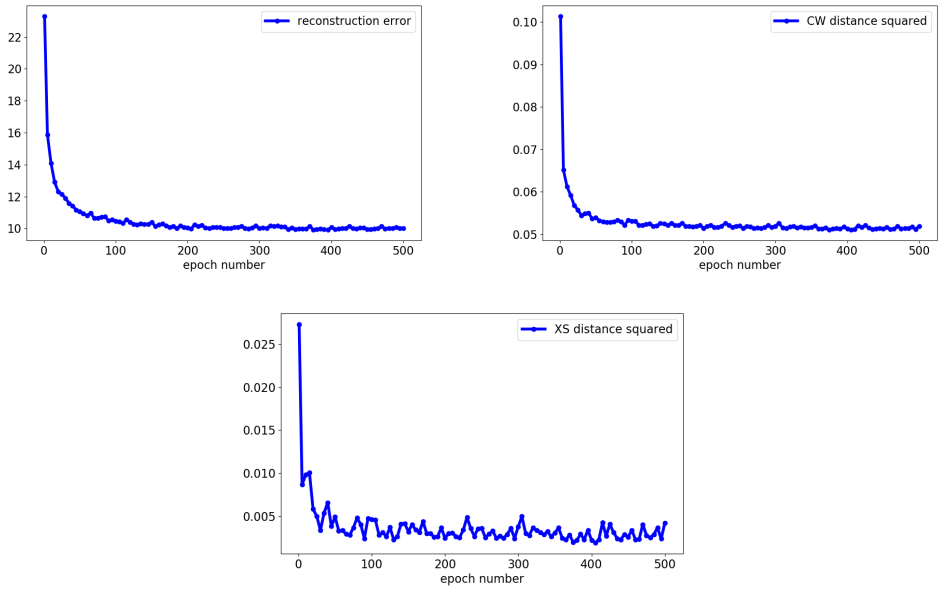

Figure 4: Convergence of the XS-VAE procedure: plot of the reconstruction error, XS and CW distances for the Fashion-MNIST dataset. XS and CW distances move in similar directions (Pearson' R correlation coefficient equals 97%) a good indication that they are equivalent for practical purposes. Also important, a reduction of the XS distance results in a reduction of the CW distance.

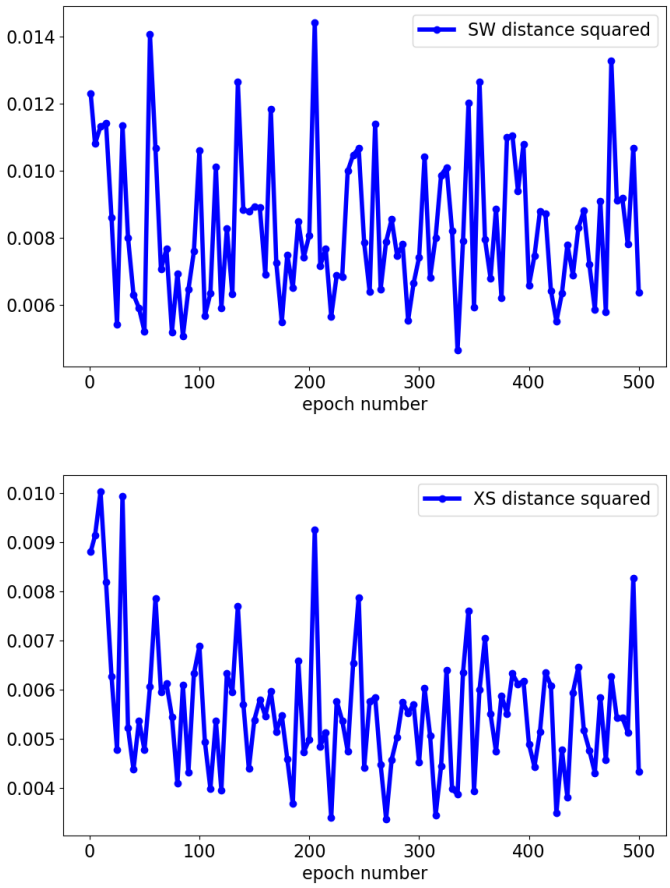

Figure 5: Convergence of the XS-VAE procedure: comparison of the SW and XS distances for the MNIST dataset. Due to convexity, the convergence of the distance is very fast and we see only the oscillation around the correct value. However this is interesting exactly because it allows to see that both distances move in a correlated manner (Pearson' R correlation coefficient is equals 64%), indicating that they are equivalent for practical purposes. Thus one can use XS distance instead of SW. Note that we do not expect a perfect linear relationship.

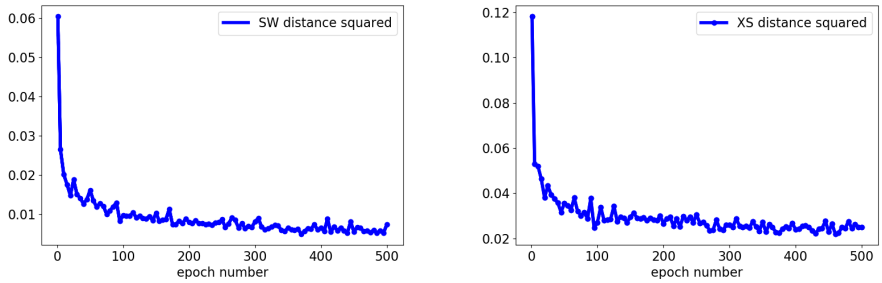

Figure 6: Convergence of the XS-VAE procedure: comparison of the SW and XS distances for the CIFAR10 dataset. Compared with MNIST dataset in Figure 5 the convergence is slower. On the other hand correlation is stronger (because the value range is larger), Pearson's R coefficient 97%.

References

[1] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, arxiv:1312.6114 (2013).

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.
URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[3] I. Tolstikhin, O. Bousquet, S. Gelly, B. Schoelkopf, Wasserstein auto-encoders, arxiv:1711.01558 (2017).

[4] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 214–223.
URL http://proceedings.mlr.press/v70/arjovsky17a.html

[5] F. John, The ultrahyperbolic differential equation with four independent variables, Duke Math. J. 4 (2) (1938) 300–322. doi:10.1215/

S0012-7094-38-00423-5.
URL https://doi.org/10.1215/S0012-7094-38-00423-5

[6] F. Santambrogio, Optimal transport for applied mathematicians. Calculus of variations, PDEs, and modeling., Vol. 87, Cham: Birkhäuser/Springer, 2015.

[7] L. Ambrosio, N. Gigli, G. Savaré, Gradient flows in metric spaces and in the space of probability measures. 2nd ed., 2nd Edition, Basel: Birkhäuser, 2008.

[8] S. Kolouri, G. K. Rohde, H. Hoffmann, Sliced Wasserstein Distance for Learning Gaussian Mixture Models, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] I. Deshpande, Z. Zhang, A. G. Schwing, Generative Modeling Using the Sliced Wasserstein Distance, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[10] S. Kolouri, P. E. Pope, C. E. Martin, G. K. Rohde, Sliced Wasserstein Auto-Encoders, in: International Conference on Learning Representations, 2019.
URL https://openreview.net/forum?id=H1xaJn05FQ

[11] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, L. V. Gool, Sliced Wasserstein generative models, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[12] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, A. G. Schwing, Max-Sliced Wasserstein Distance and Its Use for GANs, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[13] J. Tabor, S. Knop, P. Spurek, I. T. Podolak, M. Mazur, S. Jastrzkebski, Cramer-Wold AutoEncoder, CoRR abs/1805.09235. arXiv:1805.09235.
URL http://arxiv.org/abs/1805.09235

[14] G. J. Szekely, M. L. Rizzo, Energy statistics: A class of statistics based on distances, Journal of Statistical Planning and Inference 143 (8) (2013) 1249 – 1272. doi:https://doi-org-s.proxy.bu.dauphine.fr/10.1016/j.jspi.2013.03.018.
URL https://www-sciencedirect-com.proxy.bu.dauphine.fr/science/article/pii/S0378375813000633

[15] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, Y. Cheng, Sobolev GAN, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=SJA7xfb0b

[16] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, R. Munos, The Cramer Distance as a Solution to Biased Wasserstein Gradients, CoRR abs/1705.10743. arXiv:1705.10743.
URL http://arxiv.org/abs/1705.10743

[17] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, G. R. G. Lanckriet, Hilbert space embeddings and metrics on probability measures, J. Mach. Learn. Res. 11 (2010) 1517–1561.

[18] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing, The Annals of Statistics 41 (5) (2013) 2263–2291.
URL http://www.jstor.org/stable/23566550

[19] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=Hk99zCeAb

[20] R. A. Adams, J. J. F. Fournier, Sobolev spaces. 2nd ed., 2nd Edition, New York, NY: Academic Press, 2003.

[21] J. Deny, J.-L. Lions, Les espaces du type de Beppo Levi, Annales de l'Institut Fourier 5 (1954) 305–370. doi:10.5802/aif.55.
URL http://www.numdam.org/item/AIF_1954__5__305_0

[22] Peyre, Rémi, Comparison between W2 distance and H1 norm, and Localization of Wasserstein distance, ESAIM: COCV 24 (4) (2018) 1489–1501. doi:10.1051/cocv/2017050.
URL https://doi.org/10.1051/cocv/2017050

[23] F. Otto, C. Villani, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality., J. Funct. Anal. 173 (2) (2000) 361–400.

[24] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arxiv:1412.6980 (2014).

[25] M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, ninth dover printing, tenth gpo printing Edition, Dover, New York, 1964.

A Some Sobolev spaces

A.1 Spaces H^s

Recall that $L^2(\Omega)$ (also denoted L^2 when there is no ambiguity) is the space of real functions f defined on Ω such that f^2 is integrable. If $f \in L^2$ and its first derivative ∇f is also in L^2 then we say that f belongs to the Sobolev space H^1; this construction can be iterated: if f is derivable m times and all (partial) derivatives of rank m are in L^2 then $f \in H^m$. Let us take $\Omega = \mathbb{R}^N$ and recall that the Fourier transform maps the derivation operator into the multiplication by the (dual) argument. Then we can define, for any $s \geq 0$ the Sobolev space (see [20]):

$$H^s(\mathbb{R}^N) = \left\{ f \in L^2(\mathbb{R}^N) \left| \int_{\mathbb{R}^N} |\hat{f}|^2(\xi)(1+|\xi|^2)^s < \infty \right. \right\} \tag{8}$$

Here $\hat{f}$ is the Fourier transform of f. We cannot expect a probability law (such as a Dirac delta) to belong to some H^s for positive s; however it will be in some dual H^{-s} of H^s; the dual H^{-s} is a subspace of the space $S'(\mathbb{R}^N)$ of (Schwartz) distributions:

$$H^{-s}(\mathbb{R}^N) = \left\{ f \in S'(\mathbb{R}^N) \left| \int_{\mathbb{R}^N} \frac{|\hat{f}|^2(\xi)}{(1+|\xi|^2)^s} < \infty \right. \right\}. \tag{9}$$

For all $s \in \mathbb{R}$, H^s are Hilbert spaces, the squared of the norm of an element f being the integral given in the definition; H^0 reduces to L^2.

A.2 Spaces $\dot{H}^1$ and $\dot{H}^s$

For any connected domain Ω (in practice for us Ω is either an open ball in $\mathbb{R}^N$ or the whole space $\mathbb{R}^N$) let us introduce the space $BL(L^2)$ of distributions f such that $\nabla f \in L^2(\Omega)$; we also introduce the equivalence relation : $f \sim g$ if $f = g + c$ where c is a real constant. Then the quotient of $BL(L^2)$ with respect to this equivalence is denoted $\dot{H}^1$. One can prove as in [21, Corrolary 1.1] that $\dot{H}^1$ is a Hilbert space. and define its dual $\dot{H}^{-1}$ as is [22] which, for our situation, means that for any two measures μ and ν we obtain a distance d_{-1}:

$$d_{-1}(\mu,\nu)^2 = \sup \left\{ \int f(x)(\mu(dx) - \nu(dx)) \middle| \|\nabla f\|_{L^2} \leq 1 \right\}. \tag{10}$$

Note that because of the Sobolev embeddings, $\dot{H}^1$ is included in the space of continuous functions in dimension $N = 1$. Thus its dual contains any Dirac mass δ_x, $x \in \Omega$. In particular :

Proposition 2 *There exists a universal constant c_{-1} such that for $N = 1$ and $x, y \in \Omega$:*

$$d_{-1}(\delta_x, \delta_y)^2 = c_{-1}|x - y|. \tag{11}$$

Proof: We use the formulation (see [6, Section 5.5.2]): $\|\delta_x - \delta_y\|_{\dot{H}^{-1}} = \|\nabla u\|_{L^2}$ where u is the solution of the Neuman problem $\partial u/\partial n = 0$ on $\partial\Omega$ and $-\Delta u = \delta_x - \delta_y$ on Ω. But in 1D the solution is such that $|\nabla u| = |H(\cdot - \max(x, y)) - H(\cdot - \min(x, y))|$ with $H(\cdot)$ denoting the Heaviside function. □

It is remarkable to know that $\dot{H}^{-1}$ is related to the 2-Wasserstein distance W_2 through the following relation (see [22],[23, Chap. 7, formula (68)] for details): if μ is a measure with finite second order moment and ν a small variation, then formally:

$$W_2(\mu, \mu + \epsilon\nu) = |\epsilon| \cdot \|\nu\|_{\dot{H}^{-1}} + o(\epsilon). \tag{12}$$

In general, one can define for any $s \geq 0$ the homogeneous Sobolev spaces $\dot{H}^s$ as in [20].

B Analytic computations of g_{xy} and g_{xn}

We give below the formulas (and approximation results) to compute the functions g_{xy} and g_{xn} in Remark 1 for $H = \dot{H}^1$.

Proposition 3 *For $\mathcal{H} = \dot{H}^1$ up to multiplicative constants, $g_{xy}(z) = |z|$ and $g_{xn}(z) = c_{N0} + \sqrt{z^2 + c_{N1}} + O(\|x\|^4)$. More precisely:*

- *for any $x, y \in \Omega$: $d_{X\mathcal{H}}(\delta_x, \delta_y)^2 = \|x - y\|$ (up to a multiplicative constant taken as 1)*
- *Denote $N(0, I)$ the standard normal multi-variate distribution in $\mathbb{R}^N$, then*

$$d_{X\mathcal{H}}(\delta_x, N(0, I))^2 = \xi(\|x\|), \tag{13}$$

with

$$\xi(a) = (\sqrt{2} - 1)\frac{\Gamma(\frac{N+1}{2})}{\Gamma(\frac{N}{2})} - \sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \left(\frac{-a^2}{2}\right)^{k+1} \frac{1}{(k+1)!} \frac{\Gamma(\frac{N+1}{2})\Gamma(k + \frac{3}{2})}{(2k+1)\Gamma(k + 1 + \frac{N}{2})}. \tag{14}$$

Here Γ is the Euler gamma function (for instance $\Gamma(n+1) = n!$ for any integer n). In particular:

$$d_{X\mathcal{H}}(\delta_x, N(0,I))^2 = c_{N0} + \sqrt{z^2 + c_{N1}} \\ +O\left(\|x\|^4\right). \tag{15}$$

Thus the following approximation is exact up to $\|x\|^4$:

$$d_{X\mathcal{H}}\left(\frac{1}{K}\sum_{k=1}^{K}\delta_{x_k}, N(0,I)\right)^2 \simeq c_{N0} \\ +\frac{\sum_{k=1}^{K}\sqrt{\|x_k\|^2 + c_{N1}}}{K} - \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K}\|x_k - x_{k'}\|}{2K^2}, \tag{16}$$

where c_{N0}, c_{N1} are defined in (19)*.*

Proof Take $x, y \in \Omega$. When $\mathcal{H} = \dot{H}^1$ we already saw in equation (11) that the distance is translation invariant in one dimension i.e. depending only on $|\langle x - y, \theta\rangle|$. Then, by symmetry, the mean over all directions $\theta \in S^{N-1}$ of $|\langle x - y, \theta\rangle|$ is a multiple of $\|x - y\|$ (see also [14] which introduces the same distance from another point of view). The formula (11) is an application of considerations in [14, Section 4.3]. In order to derive the formula (16) it is enough to prove (15). Note first that when Z follows a N-dimensional standard normal distribution $E\|x - Z\|$ is asymptotically equal to $\|x\|$ for large values of x, which is also the case of the formula (15) (up to a constant which will not appear in the gradient). Proving the approximation (14) amounts to analyze the behavior of the function $\xi(a)$ for $a \simeq 0$. We do not take into account the universal constants (depending only on N) and consider instead the second derivative with respect to a (the first derivative is zero because the leading term is quadratic in a)

$$\xi''(0) = \sqrt{\frac{2}{\pi}}\frac{\Gamma(\frac{N+1}{2})\Gamma(\frac{3}{2})}{\Gamma(1+\frac{N}{2})}. \tag{17}$$

It is now enough to see that we also have

$$\xi(0) = (\sqrt{2} - 1)\frac{\Gamma(\frac{N+1}{2})}{\Gamma(\frac{N}{2})}, \tag{18}$$

to conclude that defining

$$c_{N0} = \xi(0) - \frac{1}{\xi''(0)}, c_{N1} = \frac{1}{\xi''(0)^2}, \tag{19}$$

we have an approximation of $\xi(a)$ exact to second order in a thus error of order $O\left(\|x\|^4\right)$ because of parity.

Remark 2 *If one uses the Stirling's formula in* (17) $\Gamma(z) = \sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z(1 + O(1/z))$ *it is possible to conclude, after straightforward computations, that* $\xi''(0) = \frac{1}{\sqrt{N}} + O(1/N)$*. This means that for some* c_1 *depending only on* N*:*

$$d_{XH}(\delta_x, N(0, I))^2 = \sqrt{\|x\|^2 + N} - c_1 + O\left(\|x\|^4 + \frac{1}{N}\right), \tag{20}$$

which is a coarser approximation and in practice gives less good results.

C Network architecture

We follow in this section the specifications in [13] and reproduce below the corresponding architectures as given in the reference:

MNIST/Fashion-MNIST (28×28 sized images):

encoder three feed-forward ReLU layers, 200 neurons each;

latent 8-dimensional;

decoder three feed-forward ReLU layers, 200 neurons each.

CIFAR-10 ($32\times$ images with 3 color layers):

encoder: four convolution layers with 2×2 filters, the second one with 2×2 strides, other non-strided (3, 32, 32, and 32 channels) with ReLU activation, 128 ReLU neurons dense layer;

latent 64-dimensional;

decoder: two dense $ReLU$ layers with 128 and 8192 neurons, two transposed-convolution layers with 2×2 filters (32 and 32 channels) and ReLU activation, a transposed convolution layer with 3×3 filter and 2×2 strides (32 channels) and ReLU activation, a transposed convolution layer with 2×2 filter (3 channels) and sigmoid activation.

The last layer returns the generated or reconstructed image.

All hyper-parameters are chosen as in the references: the loss was minimized with the Adam optimizer [24] with a learning rate of 0.001 and hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$; we used 500 epochs. The scaling parameter λ was set to 100.

D Additional details on the proofs and formulas

D.1 Additional remarks on the proofs of Propositions 2 and 3

Let us take $x, y \in \mathbb{R}$ and suppose for instance $x \leq y$. Then for any function $f \in \dot{H}$ (in particular it is absolutely continuous) we have :

$$\langle \delta_y - \delta_x, f \rangle = f(y) - f(x) = \int_x^y f'(t)dt = \int_{\mathbb{R}} 1_{[x,y]}(t) f'(t)dt. \tag{21}$$

From the definition of $\dot{H}$ it follows that the function $D_{xy} \in \dot{H}$ with $D'_{xy} = 1_{[x,y]}$ satisfies $\langle \delta_y - \delta_x, f \rangle = \langle D_{xy}, f \rangle_{\dot{H},\dot{H}}$ thus D_{xy} is the representative of $\delta_y - \delta_x \in \mathcal{H}'$ given by the Riesz theorem. In particular the norm of $\delta_y - \delta_x$ in the dual is the same as the norm of D_{xy} in $\dot{H}$. This norm (squared) equals $\int_{\mathbb{R}} |D'_{xy}|^2 dt = \int_{\mathbb{R}} 1_{[x,y]}(t)dt = y - x = |x - y|$.

Let us now take $x, y \in \Omega \subset \mathbb{R}^N$ and compute the X-ray Sovolev distance $d_{X\dot{H}}(\delta_x, \delta_y)^2$. From the previous formula:

$$d_{X\dot{H}}(\delta_x, \delta_y)^2 = \fint_{S^{N-1}} d_{\dot{H}}(\delta_{\langle \theta, x \rangle}, \delta_{\langle \theta, y \rangle})^2 d\theta. = \fint_{S^{N-1}} |\langle \theta, x - y \rangle| d\theta. \tag{22}$$

We write $x - y = \|x - y\| \theta_{xy}$ where $\theta_{xy} \in S^{N-1}$. The last integral can be written

$$\fint_{S^{N-1}} |\langle \theta, x - y \rangle| d\theta = \|x - y\| \fint_{S^{N-1}} |\langle \theta, \theta_{xy} \rangle| d\theta = c_{S^{N-1}} \|x - y\|, \tag{23}$$

where the constant $c_{S^{N-1}}$ does not depend on θ_{xy} but only on the dimension N (because the law $d\theta$ is uniform on the sphere and thus invariant to rotations).

D.2 Two more approximations for g_{xn} when $\mathcal{H} = \dot{H}$

Let us make two additional remarks concerning the computation of the distance when $\mathcal{H} = \dot{H}$. Recall that we want to compute $d_{XH}(\delta_{(x_1,...;x_N)}, N(0, I))^2$ or, equivalently, the average of $\sqrt{(x_1 - X_1)^2 + ... + (x_N - X_N)^2}$ where $x = (x_1, ..., x_N) \in \mathbb{R}^N$ and $X_1, ..., X_N$ are independent standard normal variables. Denote $Z^x = (x_1 - X_1)^2 + ... + (x_N - X_N)^2$; when $x = 0$ the random variable Z^0 has a chi-squared distribution and the average $\mathbb{E}[\sqrt{Z^0}]$ is known $\mathbb{E}[\sqrt{Z^0}] = \sqrt{2}\frac{\Gamma(\frac{N+1}{2})}{\Gamma(\frac{N}{2})}$. In general, Z^x follows a non-central χ^2 distribution with N degrees of freedom and non centrality parameter $\lambda = \|x\|^2$.

It follows that if one replaces $\mathbb{E}[\sqrt{Z^x}]$ by $\sqrt{\mathbb{E}[Z^x]} = \sqrt{\|x\|^2 + N}$ we recover the same approximation as in Remark 2. This procedure can be iterated writing

$$\mathbb{E}[\sqrt{Z^x}] = \sqrt{\mathbb{E}[Z^x]}\sqrt{1 + \frac{Z^x - \mathbb{E}[Z^x]}{\mathbb{E}[Z^x]}}, \tag{24}$$

and (formally) using the expansion (valid only for $|h| < 1$):

$$\sqrt{1+h} = 1 + \frac{h}{2} - \frac{h^2}{8} + \frac{h^3}{16} - ... \tag{25}$$

We obtain after computations:

$$\begin{aligned} &d_{XH}(\delta_{(x_1,...;x_N)}, N(0,I))^2 = \mathbb{E}[\sqrt{Z^x}] \\ &\simeq \sqrt{\|x\|^2 + N}\left(1 - \frac{2\|x\|^2 + N}{4(\|x\|^2 + N)^2} + \frac{3\|x\|^2 + N}{2(\|x\|^2 + N)^3}\right). \end{aligned} \tag{26}$$

On the other hand using the properties of the non-central chi distribution one can derive an exact formula. We know that the density of Z^x is:

$$f_{N,\lambda}(z) = \frac{e^{-\frac{z+\lambda}{2}}}{2}\left(\frac{z}{2}\right)^{N/2-1}\sum_{j=0}^{\infty}\frac{\left(\frac{\lambda z}{4}\right)^j}{j!\cdot\Gamma(j+N/2)}. \tag{27}$$

Computing $\int_0^\infty \sqrt{z} f_{N,\lambda}(z)dz$ one obtains after several straightforward algebraic manipulations (recall $\lambda = \|x\|^2$):

$$\mathbb{E}[\|x - X\|_{\mathbb{R}^N}] = \sqrt{2}\sum_{j=0}^{\infty}\frac{(\|x\|^2/2)^j}{j!\cdot e^{\|x\|^2/2}}\cdot\frac{\Gamma(j+N/2+1/2)}{\Gamma(j+N/2)}, \tag{28}$$

where X is a N-dimensional standard normal variable and $x \in \mathbb{R}^N$.

Note that the terms $p_j = \frac{(\|x\|^2/2)^j}{j!\cdot e^{\|x\|^2/2}}$ are the values of the probability mass function of the Poisson distribution of parameter $\|x\|^2/2$ (in particular they sum to 1). In principle (28) lends to direct numerical implementation but in practice, other than the care needed to evaluate the terms p_j, the sum converges very slow for large values of $\|x\|$ (hundreds up to thousands terms needed). We found that the effort to do so does not pay off (except when fast implementation in a low level programming language is used).

As a final remark note that the formula (28) allows to obtain the gradient; denote:

$$g_{xn}^N(a) := \mathbb{E}[\|x - X\|],\ x \in \mathbb{R}^N,\ \|x\| = a. \tag{29}$$

Then:

$$\frac{dg_{xn}^N(a)}{da} = a\left(g_{xn}^{N+2}(a) - g_{xn}^N(a)\right),\ \forall a \geq 0. \tag{30}$$

D.3 Additional remarks to section B

We will detail here the computations of functions g_{xy} and g_{xn} for other choices of Sobolev spaces beyond $\dot{H}$.

From formula (9) it follows that for $x, y \in \mathbb{R}$

$$\begin{aligned} \|\delta_x - \delta_y\|^2_{H^{-s}(\mathbb{R})} &= \int_{\mathbb{R}} \frac{|e^{-ix\xi} - e^{-iy\xi}|^2}{(1+|\xi|^2)^s} d\xi. \\ &= \int_{\mathbb{R}} \frac{2 - 2\cos((x-y)\xi)}{(1+|\xi|^2)^s} d\xi. \\ = \frac{\sqrt{\pi}\Gamma(s-1/2)}{2^{2s-1/2}\Gamma(s)} &- \frac{|x-y|^{s-1/2}\sqrt{\pi}}{2^{s-3/2}\Gamma(s)} K_{s-1/2}(|x-y|), \end{aligned} \tag{31}$$

where $K_s(\cdot)$ is the modified Bessel function of the second kind (see [25])

Although the formula (31) is exact, its use is somehow awkward and the analytic formula does not carry on when integrating over the unit N-dimensional sphere. An alternative formulation is possible.

Consider thus $x, y \in \Omega$. Since the distance is invariant with respect to rotations one may suppose that $x - y = \|x - y\| \cdot (1, 0, ..., 0)$, i.e., the vector is aligned with the first axis. For $\theta = (\theta_1, ..., \theta_N) \in S^{N-1}$, the scalar product $\langle x - y, \theta \rangle$ reduces to $\|x - y\|\theta_1$.

Recall that if $X_1, ..., X_N$ are independent standard normal variables then the distribution of $\left(\frac{X_1}{\sqrt{X_1^2+...+X_N^2}}, ..., \frac{X_N}{\sqrt{X_1^2+...+X_N^2}} \right)$ is the uniform law on the unit sphere in N dimensions.

In particular since $X_2, ..., X_N$ are independent standard normals, $Y = X_2 + ... + X_N$ has chi-squared distribution with $N - 1$ degrees of freedom, i.e., with probability density $\rho_N(y) = \frac{y^{N/2-3/2}e^{-y/2}}{2^{(N-1)/2}\Gamma((N-1)/2)}$. Then the distance satisfies

$$d_{XH^s}(\delta_x, \delta_y)^2 = g^{H^s}_{xy}(\|x - y\|), \tag{32}$$

where the function $g^{H^s}_{xy}$ is defined as

$$g^{H^s}_{xy}(a) = \int_{\mathbb{R}^+} \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{2 - 2\cos\left(\frac{az\xi}{\sqrt{z^2+y}}\right)}{(1+|\xi|^2)^s} p_n(z)\rho_N(y) d\xi dz dy. \tag{33}$$

Here $p_n(\cdot)$ is the density of the standard normal. This function can be accurately computed by quadrature (Gauss-Laguerre quadrature for y, Gauss-Hermite quadrature for z).

In particular from equation (33) one obtains:

$$g_{xy}^{H^s}(a) = g_{xy}^{H^s}(-a), \forall a \in \mathbb{R}, \tag{34}$$

$$g_{xy}^{H^s}(0) = 0. \tag{35}$$

Derivating once the integral one obtains a function which is odd with respect to a:

$$\frac{dg_{xy}^{H^s}(a)}{da} = \int_{\mathbb{R}^2\times\mathbb{R}^+} \frac{2z\xi \sin\left(\frac{az\xi}{\sqrt{z^2+y}}\right)}{(1+|\xi|^2)^s\sqrt{z^2+y}} p_n(z)\rho_N(y)d\xi dzdy, \tag{36}$$

and thus

$$\frac{dg_{xy}^{H^s}}{da}(0) = 0. \tag{37}$$

Moreover:

$$\frac{d^2g_{xy}^{H^s}(a)}{da^2} = \int_{\mathbb{R}^2\times\mathbb{R}^+} \frac{2z^2\xi^2 \cos\left(\frac{az\xi}{\sqrt{z^2+y}}\right)}{(1+|\xi|^2)^s(z^2+y)} p_n(z)\rho_N(y)d\xi dzdy. \tag{38}$$

In particular

$$\begin{aligned}\frac{d^2g_{xy}^{H^s}(a)}{da^2}\Big|_{a=0} &= \int_{\mathbb{R}^2\times\mathbb{R}^+} \frac{2z^2\xi^2 p_n(z)\rho_N(y)d\xi dzdy}{(1+|\xi|^2)^s(z^2+y)} \\ &= 2\int_{\mathbb{R}} \frac{\xi^2}{(1+|\xi|^2)^s}d\xi \int_{\mathbb{R}\times\mathbb{R}^+} \frac{z^2 p_n(z)\rho_N(y)dzdy}{z^2+y}.\end{aligned} \tag{39}$$

Recall now that in the last integral the variable y stands for the sum of $N-1$ squared standard normal variables i.e., the integral is nothing else than $\mathbb{E}\frac{X_1^2}{X_1^2+\ldots+X_N^2}$ which by symmetry equals $1/N$. On the other hand the first integral equals $\int_{\mathbb{R}} \frac{\xi^2}{(1+\xi^2)^s}d\xi = \frac{\sqrt{\pi}\Gamma(s-3/2)}{\Gamma(s)}$, which allows to write

$$\frac{d^2g_{xy}^{H^s}(a)}{da^2}\Big|_{a=0} = \frac{2\sqrt{\pi}\Gamma(s-3/2)}{N\Gamma(s)}. \tag{40}$$

On the other hand, for $a \to \infty$ one can have the following intuition (that can be made precise using an asymptotic expansion for the Bessel function $K_\cdot(\cdot)$ in formula (31)): the quantity $\cos\left(\frac{az\xi}{\sqrt{z^2+y}}\right)$ will oscillate rapidly around its mean value of zero. Thus in the limit $a \to \infty$ the "$\cos(\cdot)$" part will average out and only the first part remains, thus

$$\lim_{a\to\infty} g_{xy}^{H^s}(a) = \frac{2\sqrt{\pi}\Gamma(s-1/2)}{\Gamma(s)}. \tag{41}$$

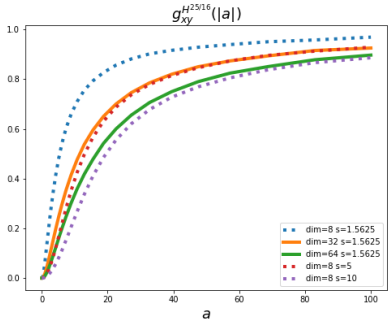

Figure 7: The function $g_{xy}^{H^s}(|a|)$ for different choices of dimension N and regularity parameter s.

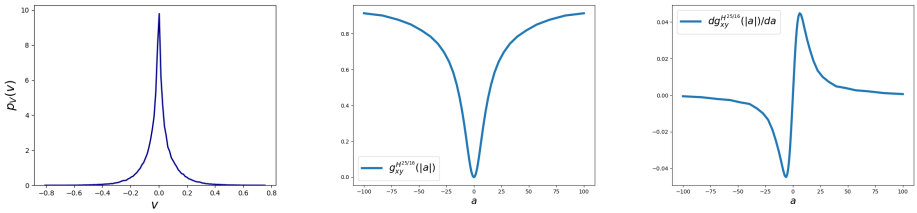

Figure 8: **Left:** The density $p_V(v)$ of V for $N = 8$ and $s = 25/16$ (s was chosen so that $s > 3/2$). **Middle:** The function $g_{xy}^{H^s}(|a|)$ computed using φ_V. **Right:** The gradient of $g_{xy}^{H^s}(|a|)$. For graphical convenience we plotted also the functions for $a < 0$ by symmetry.

D.4 Further simplifications

Consider now the real variable $U = \frac{X_1}{\sqrt{X_1^2+...+X_N^2}}$ with values in $[-1, 1]$, where $X_1, ..., X_N$ are independent standard normal variables. Then U, as real variable has a density $p_U(u)$, $u \in [-1, 1]$ depending only on N. This allows to write a simplified version of the previous formulas, for instance instead of (33) we have:

$$g_{xy}^{H^s}(a) = \int_{\mathbb{R}} \int_{[-1,1]} \frac{2 - 2\cos(au\xi)}{(1 + |\xi|^2)^s} p_U(u) d\xi du. \tag{42}$$

This construction can be iterated once more. Recall that, up to constants, if T is a random variable with Student-t distribution of parameter $2s - 1$ then $T/\sqrt{2s-1}$ has a density proportional to $\frac{1}{(1+x^2)^s}$. Consider the variable $V = UT/\sqrt{2s-1}$. Then, if we denote $p_V(v)$ the density of V (depending only on N and s and that can be precomputed), then

$$g_{xy}^{H^s}(a) = \frac{2\sqrt{\pi}\Gamma(s-1/2)}{\Gamma(s)} - 2\int_{\mathbb{R}} \cos(av)\, p_v(v)dv. \tag{43}$$

We recognize in the last term the real part of the characteristic function of the variable V.

To summarize, let: X_1 be a normal variable, Y a chi-squared (parameter $N-1$) variable and T a random variable with Student-t distribution of parameter $2s-1$; suppose all variables are independent. Then define $V = \frac{X_1 T}{\sqrt{(2s-1)\cdot(X_1^2+Y)}}$ and let φ_V be the characteristic function of V. Then, up to a multiplicative constant:

$$g_{xy}^{H^s}(a) = 1 - Re(\varphi_V(a)), \tag{44}$$

with "Re" denoting the real part of a complex number. We illustrate in figures 8 and 7 the typical behavior of the density $p_V(v)$ of V and of the distance $g_{xy}^{H^s}(a)$.