A Proof of things

We first recall notions that will be needed. We then give the proof of our main theoretical result.

A.1 Wasserstein distance and Lipschitz functions

In the following, all the metric spaces considered will be subsets of normed vector spaces, with the metric on the subset induced by the norm.

First, one recalls some definitions.

Definition 1 (transference plan). *Let $(X, \mathbb{P}_X)$ and $(Y, \mathbb{P}_Y)$ be two probability spaces. A* transference plan *γ is a measure on $X \times Y$ such that :*

$$\int_{A \times Y} d\gamma = \mathbb{P}_X(A),$$

and,

$$\int_{X \times B} d\gamma = \mathbb{P}_Y(B).$$

*$\mathbb{P}_X$ and $\mathbb{P}_Y$ are called the **marginals** of γ. The set of transference plans with marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$ is denoted by $\Pi(\mathbb{P}_X, \mathbb{P}_Y)$.*

Definition 2 (p-Wasserstein distance). *Let $(X, \|.\|)$ be a metric space and $p \in [1, +\infty)$. For two probability measure $\mathbb{P}_1$, $\mathbb{P}_2$ on X, the* p-Wasserstein distance *between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined by the following*

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \left(\inf_{\gamma \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_{(x,y)\sim\gamma} \|x - y\|^p \right)^{\frac{1}{p}}.$$

In this paper, we used the notation $W(\mathbb{P}_1, \mathbb{P}_2)$ instead of $W_1(\mathbb{P}_1, \mathbb{P}_2)$.

Definition 3 (Lipschitz function). *Let $\phi : X \to Y$ be a map between metric spaces X and Y. It is called a* C-Lipschitz *function if there exists a constant C such that :*

$$\forall x \text{ and } y \in X, \|\phi(x) - \phi(y)\|_Y \leq C\|x - y\|_X.$$

Lemma 1 *Let $\phi : X \to Y$ be a locally Lipschitz map, with X compact, then there exists a constant C such that*

$$W_Y(\phi_\sharp \mu, \phi_\sharp \nu) \leq C W_X(\mu, \nu).$$

Proof. Let γ be a transference plan realising $W_X(\mu, \nu)$. Define $\gamma' := (\phi \times \phi)_\sharp \gamma$. One can check that γ' defines a transference plan between $\phi_\sharp \mu$ and $\phi_\sharp \nu$. There-

fore, one has the following relation

$$\begin{aligned} W_Y(\phi_\sharp\mu, \phi_\sharp\nu) &\leq \int \|x-y\| d\gamma'(x,y) \\ &= \int \|\phi(x)-\phi(y)\| d\gamma(x,y) \\ &\leq \int C\|x-y\| d\gamma(x,y) \\ &= C W_X(\mu,\nu), \end{aligned}$$

where the first inequality comes from the fact that γ' is a transference plan, the first equality from the definition of the push forward of a measure by a map (recalled in section 2), the last inequality from lemma 2, and the last equality from the choice of γ.

Lemma 2 *Let $\phi : X \to Y$ be a locally Lipschitz map, and X a compact metric space. Then there exists C such that ϕ is a C-Lipschitz function.*

Proof. By definition of a locally Lipschitz map, for all x in X, there exists U_x a neighbourhood of x and a constant C_x such that ϕ is C_x-Lipschitz on U_x.
So $\bigcup_{x\in X} U_x$ is a cover of X. Since X is compact, there exists a finite set I such that $\bigcup_{i\in I} U_i$ is a cover of X.
One can check that ϕ is C-Lipschitz on X, with $C := \max_{i\in I}(C_{x_i})$.

A.2 Proof of theorem 1

We can finally turn to the proof of theorem 1 :

Theorem 1. *There exist two positive constants a and b such that*

$$\begin{aligned} W(\mathbb{P}_{\mathcal{D}'}, \mathbb{P}'_\theta) \leq a\, W(\mathbb{P}_{\mathcal{D}}, \mathbb{P}_{\mathcal{D}'}) + W(\mathbb{P}_{\mathcal{D}}, AE(\mathbb{P}_{\mathcal{D}})) \\ + b\, W(c_{1\sharp}\mathbb{P}_{\mathcal{D}'}, \mathbb{P}'^0_\theta). \end{aligned}$$

Proof. From the triangle inequality property of the Wasserstein metric and the definition of $\mathbb{P}'_\theta$, one has :

$$W(\mathbb{P}_{\mathcal{D}'}, \mathbb{P}'_\theta) \leq W(\mathbb{P}_{\mathcal{D}'}, AE(\mathbb{P}_{\mathcal{D}'})) + W(AE(\mathbb{P}_{\mathcal{D}'}), g_{1\sharp}\mathbb{P}'^0_\theta).$$

One concludes with lemma 3 and lemma 1 with $\phi = g_1$.

Lemma 3 *There exist a positive constant a such that*

$$W(\mathbb{P}_{\mathcal{D}'}, AE(\mathbb{P}_{\mathcal{D}'})) \leq aW(\mathbb{P}_{\mathcal{D}}, \mathbb{P}_{\mathcal{D}'}) + W(\mathbb{P}_{\mathcal{D}}, AE(\mathbb{P}_{\mathcal{D}})).$$

Proof. Applying twice the triangle inequality, one has :

$$\begin{aligned} W(\mathbb{P}_{\mathcal{D}'}, AE(\mathbb{P}_{\mathcal{D}'})) \leq W(\mathbb{P}_{\mathcal{D}'}, \mathbb{P}_{\mathcal{D}}) + W(\mathbb{P}_{\mathcal{D}}, AE(\mathbb{P}_{\mathcal{D}})) \\ + W(AE(\mathbb{P}_{\mathcal{D}}), AE(\mathbb{P}_{\mathcal{D}'})). \end{aligned}$$

One concludes with lemma 1 with $\phi = g_1 \circ c_1$.

Remark 1 *It is important to remark that in order to be able to apply lemma 1 in the proof of theorem 1, one needs the assumption that $\mathbb{P}'^0_\theta$ and $c_{1\sharp}\mathbb{P}_{\mathcal{D}'}$ have compact support. But as χ is itself compact, this is not a problem for $c_{1\sharp}\mathbb{P}_{\mathcal{D}'}$ since the image of a compact χ by a continuous function c_1 is compact. However, the compacity of the support of $\mathbb{P}'^0_\theta$ is not a priori granted. An easy fix is to choose a prior $\mathbb{P}_Z$ with compact support. Therefore, we choose this setting in our applications.*

Remark 2 *Our proof of theorem 1 implicitly assumed that neural networks are locally Lipschitz maps (see lemmata 1 and 3). This assumption is justified by the following lemma.*

Lemma 4 *Let $g : Z \to X$ be a neural network and $\mathbb{P}_Z$ a prior over Z such that $\mathbb{E}_{z\sim\mathbb{P}_Z}(\|z\|) < \infty$ (such as Gaussian) then g is locally Lipschitz and $\mathbb{E}_{z\sim\mathbb{P}_Z}(L_z) < \infty$, where L_z are the local Lipschitz constants.*

Proof. See Corollary 1. of [1]

A.3 Application to Wasserstein autoencoders.

The main theorem of [16], theorem 3.1, guarantees the convergence of a Wasserstein autoencoder (WAE). We recall this theorem and show that it is a direct consequence of our theorem 1.

Theorem 2

$$W(\mathbb{P}_{\mathcal{D}}, g_{1\sharp}\mathbb{P}_{\mathcal{Z}}) \leq W(\mathbb{P}_{\mathcal{D}}, AE(\mathbb{P}_{\mathcal{D}})) \tag{4}$$

$$+\ bW(c_{1\sharp}\mathbb{P}_{\mathcal{D}}, \mathbb{P}_{\mathcal{Z}}). \tag{5}$$

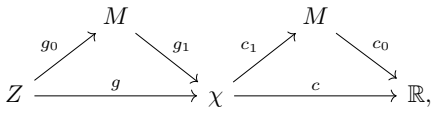
Proof. Since WAEs do not involve transfer, one has $\mathcal{D} = \mathcal{D}'$, i.e. $\mathbb{P}_{\mathcal{D}} = \mathbb{P}_{\mathcal{D}'}$. Then it suffices to replace $\mathbb{P}'^0_\theta$ by $\mathbb{P}_{\mathcal{Z}}$ and $\mathbb{P}'_\theta$ becomes $g_{1\sharp}\mathbb{P}_{\mathcal{Z}}$.

Remark 3 *Our proof of theorem 1 is very similar to the proof of theorem 2 given in [16]. Therefore, our contribution here consists rather in finding a versatile statement that applies to both problems (transfer and WAE) than in the originality of the tools used in the proofs.*

Remark 4 *When one restricts our approach to the case when $\mathcal{D} = \mathcal{D}'$, it does not coincide with WAE. Indeed, with the notations of our paper, WAE work with a fixed prior $\mathbb{P}_M$ on M that one tries to approximate by $c_{1\sharp}\mathbb{P}_{\mathcal{D}}$, while constraining c_1 to be a right inverse (in measure) of g_1, and $g_{1\sharp}\mathbb{P}_M$ to approximate (in measure) $\mathbb{P}_{\mathcal{D}}$. On the other hand, our approach involves an extra auxiliary latent space Z. Therefore we can consider $g_{0\sharp}\mathbb{P}_Z$ as a replacement of $\mathbb{P}_M$. Via the flexibility of the learnable weights of g_0, we use $g_{0\sharp}\mathbb{P}_Z$ to approximate $c_{1\sharp}\mathbb{P}_{\mathcal{D}}$, instead of using $c_{1\sharp}\mathbb{P}_{\mathcal{D}}$ to approximate $\mathbb{P}_M$ as in [14]. This is fundamental, because in a setting where $\mathcal{D} \neq \mathcal{D}'$, this decoupling permits to train c_1 and g_1 on $\mathcal{D}$ and c_0 and g_0 on $c_1(\mathcal{D}')$, enabling us to do transfer.*

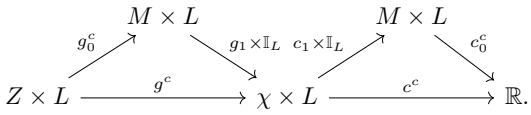
B Mind2Mind conditional GANs

As suggested to us by L. Cetinsoy, the Mind2Mind approach also applies to conditional GANs. However, one needs to implement the following modifications : replace M by $M \times L$ and Z by $Z \times L$ in the diagram



(6)

where L stands for the space of conditions, in order to get



(7)

Here, (g_0^c, c_0^c) and (g^c, c^c) are conditional GANs, with the generators of the form $g_0^c(z,l) = (m(z,l),l)$ and $g^c(z,l) = (x(z,l),l)$. The autoencoder $(c_1 \times \mathbb{I}_L, g_1 \times \mathbb{I}_L)$ can be trivially deduced from an autoencoder (c_1, g_1) via the formulas $c_1 \times \mathbb{I}_L(x,l) := (c_1(x),l)$ and $g_1 \times \mathbb{I}_L(m,l) := (c_1(m),l)$.
In practice, the algorithm 1 becomes a classical conditional GAN algorithm :

Algorithm 2 Conditional-MindGAN transfer learning.

Require: (c_1, g_1), an autoencoder trained on a source dataset $\mathcal{D}$, α, the learning rate, b, the batch size, n, the number of iterations of the critic per generator iteration, $\mathcal{D}' \subset \chi \times L$, a dataset with conditions, φ' and θ' the initial parameters of the critic c_0^c and of the generator g_0^c.
Compute $(c_1 \times \mathbb{I}_L)(\mathcal{D}')$.
while θ' has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{(m^{(i)}, l^{(i)})\}_{i=1}^b \sim (c_1 \times \mathbb{I}_L)_\sharp \mathbb{P}_{\mathcal{D}'}$ a batch from $(c_1 \times \mathbb{I}_L)(\mathcal{D}')$.
    Sample $\{(z^{(i)}, l^{(i)})\}_{i=1}^b \sim \mathbb{P}_{Z \times L}$ a batch of prior samples with conditions.
    Update c_0^c by descending L_c.
  end for
  Sample$\{(z^{(i)}, l^{(i)})\}_{i=1}^b \sim \mathbb{P}_{Z \times L}$ a batch of prior samples with conditions.
  Update g_0^c by descending $-L_g$.
end while
return $g_1 \circ g_0^c$.

C Supplementary experiments

In figure 3 we display additional samples from a MindGAN on CelebaHQ transferred from FFHQ. We also display in figure 4 the mean and standard deviation

Fig. 3: Mind2Mind on CelebaHQ transfered from FFHQ .

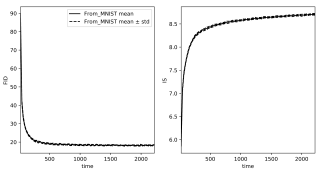


Fig. 4: Mean and standard deviation of the training of a MindGAN.

over 10 runs of the training of a MindGAN in 28×28. The source dataset in figure 5 is $\mathcal{D}' = \text{FashionMNIST}$, while in figure 6, $\mathcal{D}' = \text{MNIST}$. For comparison, we display in figure 7 samples from a vanilla WGAN.

Fig. 5: MindGAN from FashionMNIST.

Fig. 6: MindGAN from MNIST.

Fig. 7: Vanilla WGAN.