Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses

Eloi Tanguy

Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses

Eloi Tanguy[1]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

October 9, 2023

Abstract

Optimal Transport has sparked vivid interest in recent years, in particular thanks to the Wasserstein distance, which provides a geometrically sensible and intuitive way of comparing probability measures. For computational reasons, the Sliced Wasserstein (SW) distance was introduced as an alternative to the Wasserstein distance, and has seen uses for training generative Neural Networks (NNs). While convergence of Stochastic Gradient Descent (SGD) has been observed practically in such a setting, there is to our knowledge no theoretical guarantee for this observation. Leveraging recent works on convergence of SGD on non-smooth and non-convex functions by Bianchi et al. (2022), we aim to bridge that knowledge gap, and provide a realistic context under which fixed-step SGD trajectories for the SW loss on NN parameters converge. More precisely, we show that the trajectories approach the set of (sub)-gradient flow equations as the step decreases. Under stricter assumptions, we show a much stronger convergence result for noised and projected SGD schemes, namely that the long-run limits of the trajectories approach a set of generalised critical points of the loss function.

Table of Contents

1 Introduction

1.1 Optimal Transport in Machine Learning

Optimal Transport (OT) allows the comparison of measures on a metric space by generalising the use of the ground metric. Typical applications use the so-called 2-Wasserstein distance, defined as

$$\forall \mathbb{x}, \mathbb{y} \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{W}_2^2(\mathbb{x}, \mathbb{y}) := \inf_{\mathbb{\pi} \in \Pi(\mathbb{x}, \mathbb{y})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \mathrm{d}\mathbb{\pi}(x, y), \tag{W2}$$

where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ admitting a second-order moment and where $\Pi(\mathbb{x}, \mathbb{y})$ is the set of measures of $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ of first marginal $\mathbb{x}$ and second marginal $\mathbb{y}$. One may find a thorough presentation of its properties in classical monographs such as Peyré & Cuturi (2019); Santambrogio (2015); Villani (2009)

The ability to compare probability measures is useful in probability density fitting problems, which are a sub-genre of generation tasks. In this formalism, one considers a probability measure parametrised by a vector u which is designed to approach a target data distribution $\mathbb{y}$ (typically the real-world dataset). In order to determine suitable parameters, one may choose any probability discrepancy (Kullback-Leibler, Ciszar divergences, f-divergences or Maximum Mean Discrepancy (Gretton et al., 2006)), or in our case, the Wasserstein distance. In the case of Generative Adversarial Networks, the optimisation problem which trains the "Wasserstein GAN" (Arjovsky et al., 2017) stems from the Kantorovitch-Rubinstein dual expression of the 1-Wasserstein distance.

1.2 The Sliced Wasserstein Distance as an Alternative

The Wasserstein distance suffers from the curse of dimensionality, in the sense that the sample complexity for n samples in dimension d is of the order $\mathcal{O}(n^{1/d})$ (Dudley, 1969). Due to this practical limitation and to the computational cost of the Wasserstein distance, the study of cheaper alternatives has become a prominent field of research. A prominent example is the Sinkhorn Divergence introduced by Cuturi (2013), which adds an entropic regularisation term, advantageously making the problem strongly convex. Sample complexity bounds have been derived by Genevay et al. (2019), showing a convergence in $\mathcal{O}(\sqrt{n})$ with a constant depending on the regularisation factor.

Another alternative is the Sliced Wasserstein (SW) Distance introduced by Rabin et al. (2012), which consists in computing the 1D Wasserstein distances between projections of input measures, and averaging over the projections. The aforementioned projection of a measure $\mathbb{x}$ on $\mathbb{R}^d$ is done by the *push-forward* operation by the map $P_\theta : x \longmapsto \theta^\top x$. Formally, $P_\theta \# \mathbb{x}$ is the measure on $\mathbb{R}$ such that for any Borel set $B \subset \mathbb{R}$, $P_\theta \# \mathbb{x}(B) = \mathbb{x}(P_\theta^{-1}(B))$. Once the measures are projected onto a line $\mathbb{R}\theta$, the computation of the Wasserstein distance becomes substantially simpler numerically. We illustrate this fact in the discrete case, which arises in practical optimisation settings. Let two discrete measures on $\mathbb{R}^d$: $\mathbb{\delta}_X := \frac{1}{n}\sum_k \mathbb{\delta}_{x_k}$, $\mathbb{\delta}_Y := \frac{1}{n}\sum_k \mathbb{\delta}_{y_k}$ with supports $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n) \in \mathbb{R}^{n \times d}$. Their push-forwards by P_θ are simply computed by the formula $P_\theta \# \mathbb{\delta}_X = \frac{1}{n}\sum_k \mathbb{\delta}_{P_\theta(x_k)}$, and the 2-Wasserstein distance between their projections can be computed by sorting their supports: let σ a permutation sorting $(\theta^\top x_1, \cdots, \theta^\top x_n)$, and τ a permutation sorting $(\theta^\top y_1, \cdots, \theta^\top y_n)$, one has the simple expression

$$\mathrm{W}_2^2(P_\theta \# \mathbb{\delta}_X, P_\theta \# \mathbb{\delta}_Y) = \frac{1}{n}\sum_{k=1}^{n} (\theta^\top x_{\sigma(k)} - \theta^\top y_{\tau(k)})^2. \tag{1}$$

The SW distance is the expectation of this quantity with respect to $\theta \sim \mathbb{\sigma}$, i.e. uniform on the sphere: $\mathrm{SW}_2^2(\mathbb{\delta}_X, \mathbb{\delta}_Y) = \mathbb{E}_{\theta \sim \mathbb{\sigma}}\left[\mathrm{W}_2^2(P_\theta \# \mathbb{\delta}_X, P_\theta \# \mathbb{\delta}_Y)\right]$. The 2-SW distance is also defined more generally between two measures $\mathbb{x}, \mathbb{y} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\mathrm{SW}_2^2(\mathbb{x}, \mathbb{y}) := \int_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \mathbb{x}, P_\theta \# \mathbb{y}) \mathrm{d}\mathbb{\sigma}(\theta). \tag{SW}$$

In addition to its computational accessibility, the SW distance enjoys a dimension-free sample complexity (Nadjahi et al., 2020). Additional statistical, computational and robustness properties of SW have been explored by Nietert et al. (2022). Moreover, central-limit results have been shown by Xu & Huang (2022) for 1-SW and the 1-max-SW distance (a variant of SW introduced by Deshpande et al. (2019)), and related work by Xi & Niles-Weed (2022) shows the convergence of the sliced error process $\theta \longmapsto \sqrt{n}\left(\mathrm{W}_p^p(P_\theta \# \mathbb{\alpha}_X, P_\theta \# \mathbb{\alpha}_Y) - \mathrm{W}_p^p(P_\theta \# \alpha, P_\theta \# \beta)\right)$, where the samples $X \sim \alpha^{\otimes n}$ and $Y \sim \beta^{\otimes n}$ are drawn for each θ. Another salient field of research for SW is its metric properties, and while it has been shown to be weaker than the Wasserstein distance in general by Bonnotte (2013), and metric comparisons with Wasserstein and max-SW have been undergone by Bayraktar & Guo (2021) and Paty & Cuturi (2019).

1.3 Related Works

Our subject of interest is the theoretical properties of SW as a loss for implicit generative modelling, which leads to minimising $\mathrm{SW}_2^2(T_u \# \alpha, \beta)$ in the parameters u, where β is the target distribution, and $T_u \# \alpha$ is the image by the NN[1] of α, a low-dimensional input distribution (often chosen as Gaussian or uniform noise). In order to train a NN in this manner, at each iteration one draws n samples from α and β (denoted $\mathbb{\alpha}_X$ and $\mathbb{\alpha}_Y$ as discrete measures with n points), as well as a projection θ (or a batch of projections) and performs an SGD step on the sample loss

$$\mathcal{L}(u) = \mathrm{SW}_2^2(P_\theta \# T_u \# \mathbb{\alpha}_X, P_\theta \# \mathbb{\alpha}_Y) = \frac{1}{n}\sum_{k=1}^{n} (\theta^\top T_u(x_{\sigma(k)}) - \theta^\top y_{\tau(k)})^2. \tag{2}$$

Taking the expectation of this loss over the samples yields the minibatch Sliced-Wasserstein discrepancy, a member of the minibatch variants of the OT distances, introduced formally by Fatras et al. Fatras et al. (2021). The framework (2) fits several Machine Learning applications, for instance, Deshpande et al. (2018) trains GANs and auto-encoders with this method, and Wu et al. (2019) consider related dual formulations. Other examples within this formalism include the synthesis of images by minimising the SW distance between features of the optimised image and a target image, as done by Heitz et al. (2021) for textures with neural features, and by Tartavel et al. (2016) with wavelet features (amongst other methods).

The general study of convergence of SGD in the context of non-smooth, non-convex functions (as is the case of $\mathcal{L}$ from (2)) is an active field of research: Majewski et al. (2018) and Davis et al. (2020) show the convergence of diminishing-step SGD under regularity constraints, while Bolte & Pauwels (2021) leverage conservative field theory to show convergence results for training with back-propagation. Finally, the recent work by Bianchi et al. (2022) shows the convergence of fixed-step SGD schemes on a general function F under weaker regularity assumptions.

More specifically, the study of convergence for OT-based generative NNs has been tackled by Fatras et al. (2021), who prove strong convergence results for minibatch variants of classical OT distances, namely the Wasserstein distance, the Sinkhorn Divergence and the Gromov Wasserstein distance (another OT variant introduced by Mémoli (2011)). A related study on GANs by Huang et al. (2023) derive optimisation properties for one layer and one dimensional Wasserstein-GANs and generalise to higher dimensions by turning to SW-GANs. Another work by Bréchet et al. (2023) focuses on the theoretical properties of linear NNs trained with the Bures-Wasserstein loss (introduced by Bures (1969); see also (Bhatia et al., 2017) for reference on this metric). Finally, the regularity and optimisation properties of the simpler energy $\mathrm{SW}_2^2(\mathbb{\alpha}_X, \mathbb{\alpha}_Y)$ have been studied by Tanguy et al. (2023).

In practice, it has been observed that SGD in such settings always converges (in the loose numerical sense, see (Deshpande et al., 2018), Section 5, or (Heitz et al., 2021), Figure 3), yet this property is not known theoretically. The aim of this work is to bridge the gap between theory and practical

[1] Similarly to the 1D case, $T_u \# \alpha$ is the push-forward measure of α by T_u, i.e. the law of $T_u(x)$ when $x \sim \alpha$.

observation by proving convergence results for SGD on (minibatch) Sliced Wasserstein generative losses of the form $F(u) = \mathbb{E}_{X\sim\mathbb{x}^{\otimes n},Y\sim\mathbb{y}^{\otimes n}}\mathrm{SW}_2^2(T_u\#\mathbb{O}X,\mathbb{O}Y)$.

1.4 Contributions

Convergence of Interpolated SGD Under Practical Assumptions Under practically realistic assumptions, we prove in Theorem 1 that piecewise affine interpolations (defined in Equation (10)) of constant-step SGD schemes on $u \longmapsto F(u)$ (formalised in Equation (7)) converge towards the set of sub-gradient flow solutions (see Equation (9)) as the gradient step decreases. This results signifies that with very small learning rates, SGD trajectories will be close to sub-gradient flows, which themselves converge to critical points of F (omitting serious technicalities).

The assumptions for this result are practically reasonable: the input measure $\mathbb{x}$ and the true data measure $\mathbb{y}$ are assumed to be compactly supported. As for the network $(u,x) \longmapsto T(u,x)$, we assume that for a fixed datum x, $T(\cdot,x)$ is piecewise $\mathcal{C}^2$-smooth and that it is Lipschitz jointly in both variables. We require additional assumptions on T which are more costly, but are verified as long as T is a NN composed of typical activations and linear units, with the constraint that the parameters u and data x stay both stay within a fixed bounded domains. We discuss a class of neural networks that satisfy all of the assumptions of the paper in the Appendix (Section D). Furthermore, this result can be extended to other orders $p \neq 2$ of SW: we present the tools for this generalisation in Section E.

Stronger Convergence Under Stricter Assumptions In order to obtain a stronger convergence result, we consider a variant of SGD where each iteration receives an additive noise (scaled by the learning rate) which allows for better space exploration, and where each iteration is projected on a ball $B(0,r)$ in order to ensure boundedness. This alternative SGD scheme remains within the realm of practical applications, and we show in Theorem 2 that long-run limits of such trajectories converge towards a set of generalised critical points of F, as the gradient step approaches 0. This result is substantially stronger, and can serve as an explanation of the convergence of practical SGD trajectories, specifically towards a set of critical points which amounts to the stationary points of the energy (barring theoretical technicalities).

Unfortunately, we require additional assumptions in order to obtain this stronger convergence result, the most important of which is that the input data measure $\mathbb{x}$ and the dataset measure $\mathbb{y}$ are discrete. For the latter, this is always the case in practice, however the former assumption is more problematic, since it is common to envision generative NNs as taking an argument from a continuous space (the input is often Gaussian of Uniform noise), thus a discrete setting is a substantial theoretical drawback. For practical concerns, one may argue that the discrete $\mathbb{x}$ can have an arbitrary fixed amount of points, and leverage strong sample complexity results to ascertain that the discretisation is not costly if the number of samples is large enough.

2 Stochastic Gradient Descent with SW as Loss

Training Sliced-Wasserstein generative models consists in training a neural network

$$T : \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ (u,x) & \longmapsto & T_u(x) := T(u,x) \end{array} \right. \tag{3}$$

by minimising the SW minibatch loss $u \longmapsto \mathbb{E}_{X\sim\mathbb{x}^{\otimes n},Y\sim\mathbb{y}^{\otimes n}}\left[\mathrm{SW}_2^2(T_u\#\mathbb{O}X,\mathbb{O}Y)\right]$ through Stochastic Gradient Descent (as described in Algorithm 1). The probability distribution $\mathbb{x} \in \mathcal{P}_2(\mathbb{R}^{d_x})$ is the law of the input of the generator $T(u,\cdot)$. The distribution $\mathbb{y} \in \mathcal{P}_2(\mathbb{R}^{d_y})$ is the data distribution, which T aims to simulate. Finally, $\mathbb{o}$ will denote the uniform measure on the unit sphere of $\mathbb{R}^{d_y}$, denoted by $\mathbb{S}^{d_y-1}$. Given a list of points $X = (x_1,\cdots,x_n) \in \mathbb{R}^{n\times d_x}$, denote the associated discrete uniform measure $\mathbb{O}X := \frac{1}{n}\sum_i \mathbb{O}_{x_i}$. By abuse of notation, we write $T_u(X) := (T_u(x_1),\cdots,T_u(x_n)) \in \mathbb{R}^{n\times d_y}$. The reader may find a summary of this paper's notations in Table 1.

Algorithm 1: Training a NN on the SW loss with Stochastic Gradient Descent

Data: Learning rate $\alpha > 0$, probability distributions $\mathbb{x} \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\mathbb{y} \in \mathcal{P}_2(\mathbb{R}^{d_y})$.

Initialisation: Draw $u^{(0)} \in \mathbb{R}^{d_u}$;

for $t \in [\![0, T_{\max} - 1]\!]$ **do**

  Draw $\theta^{(t+1)} \sim \mathbb{o}$, $X^{(t+1)} \sim \mathbb{x}^{\otimes n}$ $Y^{(t+1)} \sim \mathbb{y}^{\otimes n}$. SGD update:

$$u^{(t+1)} = u^{(t)} - \alpha \left[\frac{\partial}{\partial u} \mathrm{W}_2^2(P_{\theta^{(t+1)}} \# T_u \# \mathbb{O}_{X^{(t+1)}}, P_{\theta^{(t+1)}} \# \mathbb{O}_{Y^{(t+1)}}) \right]_{u=u^{(t)}}$$

end

In the following, we will apply results from (Bianchi et al., 2022), and we pave the way to the application of these results by presenting their theoretical framework. Consider a sample loss function $f : \mathbb{R}^{d_u} \times \mathcal{Z} \longrightarrow \mathbb{R}$ that is locally Lipschitz in the first variable, and $\mathbb{z}$ a probability measure on $\mathcal{Z} \subset \mathbb{R}^d$ which is the law of the samples drawn at each SGD iteration. Consider $\varphi : \mathbb{R}^{d_u} \times \mathcal{Z} \longrightarrow \mathbb{R}^{d_u}$ an *almost-everywhere gradient* of f, which is to say that for almost every $(u, z) \in \mathbb{R}^{d_u} \times \mathcal{Z}$, $\varphi(u, z) = \partial_u f(u, z)$ (since each $f(\cdot, z)$ is locally Lipschitz, it is differentiable almost-everywhere by Rademacher's theorem). The complete loss function is the expectation of the sample loss, $F := u \longrightarrow \int_{\mathcal{Z}} f(u, z) \mathrm{d}\mathbb{z}(z)$. An SGD trajectory of step $\alpha > 0$ for F is a sequence $(u^{(t)}) \in (\mathbb{R}^{d_u})^{\mathbb{N}}$ of the form:

$$u^{(t+1)} = u^{(t)} - \alpha \varphi(u^{(t)}, z^{(t+1)}), \quad \left(u^{(0)}, (z^{(t)})_{t \in \mathbb{N}}\right) \sim \mathbb{u}_0 \otimes \mathbb{z}^{\otimes \mathbb{N}},$$

where $\mathbb{u}_0$ is the distribution of the initial position $u^{(0)}$. Within this framework, we define an SGD scheme described by Algorithm 1, with $\mathbb{z} := \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \mathbb{o}$ and the minibatch SW sample loss

$$f := \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{n \times d_x} \times \mathbb{R}^{n \times d_y} \times \mathbb{S}^{d_y - 1} & \longrightarrow & \mathbb{R}^{d_y} \\ (u, X, Y, \theta) & \longmapsto & \mathrm{W}_2^2(P_\theta \# T_u \# \mathbb{O}_X, P_\theta \# \mathbb{O}_Y) \end{array} \right. . \tag{4}$$

With this definition for f, we have

$$F(u) = \mathbb{E}_{(X,Y,\theta) \sim \mathbb{z}} \left[f(u, X, Y, \theta) \right] = \mathbb{E}_{(X,Y) \sim \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n}} \left[\mathrm{SW}_2^2(T_u \# \mathbb{O}_X, \mathbb{O}_Y) \right], \tag{5}$$

thus the population loss compares the "true" data $\mathbb{y}$ with the model's generation $T_u \# \mathbb{x}$ using (minibatch) SW. We now wish to define an almost-everywhere gradient of f. To this end, notice that one may write $f(u, X, Y, \theta) = w_\theta(T(u, X), Y)$, where for $X, Y \in \mathbb{R}^{n \times d_y}$ and $\theta \in \mathbb{S}^{d_y - 1}$, $w_\theta(X, Y) := \mathrm{W}_2^2(P_\theta \# \mathbb{O}_X, P_\theta \# \mathbb{O}_Y)$. The differentiability properties of $w_\theta(\cdot, Y)$ are already known (Tanguy et al., 2023; Bonneel et al., 2015), in particular one has the following almost-everywhere gradient of $w_\theta(\cdot, Y)$:

$$\frac{\partial w_\theta}{\partial X}(X, Y) = \left(\frac{2}{n} \theta \theta^\top (x_k - y_{\sigma_\theta^{X,Y}(k)}) \right)_{k \in [\![1,n]\!]} \in \mathbb{R}^{n \times d_y},$$

where the permutation $\sigma_\theta^{X,Y} \in \mathfrak{S}_n$ is $\tau_Y^\theta \circ (\tau_X^\theta)^{-1}$, with $\tau_Y^\theta \in \mathfrak{S}_n$ being a sorting permutation of the list $(\theta^\top y_1, \cdots, \theta^\top y_n)$. The sorting permutations are chosen arbitrarily when there is ambiguity. To define an almost-everywhere gradient, we must differentiate $f(\cdot, X, Y, \theta) = u \longmapsto w_\theta(T(u, X), Y)$ for which we need regularity assumptions on T: this is the goal of Assumption 1. In the following, $\overline{A}$ denotes the topological closure of a set A, ∂A its boundary, and $\mathbb{\lambda}_{\mathbb{R}^{d_u}}$ denotes the Lebesgue measure of $\mathbb{R}^{d_u}$.

Assumption 1. *For every $x \in \mathbb{R}^{d_x}$, there exists a family of disjoint connected open sets $(\mathcal{U}_j(x))_{j \in J(x)}$ such that $\forall j \in J(x)$, $T(\cdot, x) \in \mathcal{C}^2(\mathcal{U}_j(x), \mathbb{R}^{d_y})$, $\bigcup_{j \in J(x)} \overline{\mathcal{U}_j(x)} = \mathbb{R}^{d_u}$ and $\mathbb{\lambda}_{\mathbb{R}^{d_u}}\Big(\bigcup_{j \in J(x)} \partial \mathcal{U}_j(x) \Big) = 0$.*

Note that for measure-theoretic reasons, the sets $J(x)$ are assumed countable. One may understand this assumption broadly as the neural networks T being piecewise smooth with respect to the parameters u, where the pieces depend on the input data x. In practice, Assumption 1 is an assumption

on the activation functions of the neural network. For instance, it is of course satisfied in the case of smooth activations, or in the common case of piecewise polynomial activations. We detail suitable neural networks in the Appendix (Section D).

Assumption 1 implies that given X, Y, θ fixed, $f(\cdot, X, Y, \theta)$ is differentiable almost-everywhere, and that one may define the following almost-everywhere gradient (6).

$$\varphi : \begin{cases} \mathbb{R}^{d_u} \times \mathbb{R}^{n\times d_x} \times \mathbb{R}^{n\times d_y} \times \mathbb{S}^{d_y-1} & \longrightarrow & \mathbb{R}^{d_u} \\ (u, X, Y, \theta) & \longmapsto & \displaystyle\sum_{k=1}^{n} \frac{2}{n} \left(\frac{\partial T}{\partial u}(u, x_k)\right)^\top \theta\theta^\top (T(u, x_k) - y_{\sigma_\theta^{T(u,X),Y}(k)}) \end{cases}, \tag{6}$$

where for $x \in \mathbb{R}^{d_x}$, $\dfrac{\partial T}{\partial u}(u, x) \in \mathbb{R}^{d_y \times d_u}$ denotes the matrix of the differential of $u \longmapsto T(u, x)$, which is defined for almost-every u. Given $u \in \partial \mathcal{U}_j(x)$ (a point of potential non-differentiability), take instead 0. (Any choice at such points would still define an a.e. gradient, and will make no difference).

Given a step $\alpha > 0$, and an initial position $u^{(0)} \sim \mathfrak{u}_0$, we may now define formally the following fixed-step SGD scheme for F:

$$\begin{gathered} u^{(t+1)} = u^{(t)} - \alpha \varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}), \\ \left(u^{(0)}, (X^{(t)})_{t\in\mathbb{N}}\ (Y^{(t)})_{t\in\mathbb{N}}\ (\theta^{(t)})_{t\in\mathbb{N}}\right) \sim \mathfrak{u}_0 \otimes \mathfrak{x}^{\otimes\mathbb{N}} \otimes \mathfrak{y}^{\otimes\mathbb{N}} \otimes \mathfrak{o}^{\otimes\mathbb{N}}. \end{gathered} \tag{7}$$

An important technicality that we must verify in order to apply Bianchi et al. (2022)'s results is that $u \longmapsto f(u, X, Y, \theta)$ and F are locally Lipschitz. Before proving those claims, we reproduce a useful Property from (Tanguy et al., 2023). In the following, $\|X\|_{\infty,2}$ denotes $\max_{k\in[\![1,n]\!]} \|x_k\|_2$ given $X = (x_1, \cdots, x_n) \in \mathbb{R}^{n\times d_x}$, and $B_{\mathcal{N}}(x, r)$ for $\mathcal{N}$ a norm on $\mathbb{R}^{d_x}$, $x \in \mathbb{R}^{d_x}$ and $r > 0$ shall denote the open ball of $\mathbb{R}^{d_x}$ of centre x and radius r for the norm $\mathcal{N}$ (if $\mathcal{N}$ is omitted, then B is an euclidean ball).

Proposition 1. *The $(w_\theta(\cdot, Y))_{\theta\in\mathbb{S}^{d_y-1}}$ are uniformly locally Lipschitz (Tanguy et al., 2023) Prop. 2.2.1.*

Let $K_w(r, X, Y) := 4n(r + \|X\|_{\infty,2} + \|Y\|_{\infty,2})$, for $X, Y \in \mathbb{R}^{n\times d_y}$ and $r > 0$. Then $w_\theta(\cdot, Y)$ is $K_w(r, X, Y)$-Lipschitz in the neighbourhood $B_{\|\cdot\|_{\infty,2}}(X, r)$:

$$\forall Y', Y'' \in B_{\|\cdot\|_{\infty,2}}(X, r),\ \forall \theta \in \mathbb{S}^{d_y-1},\ |w_\theta(Y', Y) - w_\theta(Y'', Y)| \leq K_w(r, X, Y)\|Y' - Y''\|_{\infty,2}.$$

In order to deduce regularity results on f and F from Proposition 1, we will make the assumption that T is globally Lipschitz in (u, x). In practice, this is the case when both parameters are enforced to stay within a fixed bounded domain, for instance by multiplying a typical NN with the indicator of such a set. We present this in detail in the Appendix (Section D).

Assumption 2. *There exists $L > 0$ such that*

$$\forall (u_1, u_2, x_1, x_2) \in (\mathbb{R}^{d_u})^2 \times (\mathbb{R}^{d_x})^2,\ \|T(u_1, x_1) - T(u_2, x_2)\|_2 \leq L\left(\|u_1 - u_2\|_2 + \|x_1 - x_2\|_2\right).$$

Proposition 2. *Under Assumption 2, for $\varepsilon > 0$, $u_0 \in \mathbb{R}^{d_u}$, $X \in \mathbb{R}^{n\times d_x}$, $Y \in \mathbb{R}^{n\times d_y}$ and $\theta \in \mathbb{S}^{d_y-1}$, let $K_f(\varepsilon, u_0, X, Y) := 4Ln(\varepsilon L + \|T(u_0, X)\|_{\infty,2} + \|Y\|_{\infty,2})$. Then $f(\cdot, X, Y, \theta)$ is $K_f(\varepsilon, u_0, X, Y)$-Lipschitz in $B(u_0, \varepsilon)$:*

$$\forall u, u' \in B(u_0, \varepsilon),\ |f(u, X, Y, \theta) - f(u', X, Y, \theta)| \leq K_f(\varepsilon, u_0, X, Y)\|u - u'\|_2.$$

Proof. Let $\varepsilon > 0$, $u_0 \in \mathbb{R}^{d_u}$, $X \in \mathbb{R}^{n\times d_x}$, $Y \in \mathbb{R}^{n\times d_y}$ and $\theta \in \mathbb{S}^{d_y-1}$. Let $u, u' \in B(u_0, \varepsilon)$. Using Assumption 2, we have $T(u, X), T(u', X) \in B_{\|\cdot\|_{\infty,2}}(T(u_0, X), r)$, with $r := \varepsilon L$.

Denoting $L := L_{\overline{B}(u_0,\varepsilon),\overline{B}(0_{\mathbb{R}^{d_x}},\|X\|_{\infty,2})}$, we apply successively Proposition 1 (first inequality), then Assumption 2 (second inequality):

$$\begin{aligned}|f(u,X,Y,\theta) - f(u',X,Y,\theta)| &= |w_\theta(T(u,X),Y) - w_\theta(T(u',X),Y)| \\ &\leq K_w(r,T(u_0,X),Y)\|T(u,X) - T(u',X)\|_{\infty,2} \\ &\leq 4n(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2})L\|u-u'\|_2.\end{aligned}$$

□

Proposition 2 shows that f is locally Lipschitz in u. We now assume some conditions on the measures $\mathbb{x}$ and $\mathbb{y}$ in order to prove that F is also locally Lipschitz. Specifically, we require that the data measures $\mathbb{x}$ and $\mathbb{y}$ be supported on bounded domains, which imposes little restriction in practice.

Assumption 3. *$\mathbb{x}$ and $\mathbb{y}$ are Radon probability measures on $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$ respectively, supported by the compacts $\mathcal{X}$ and $\mathcal{Y}$ respectively. Denote $R_x := \sup_{x\in\mathcal{X}} \|x\|_2$ and $R_y := \sup_{y\in\mathcal{Y}} \|y\|_2$.*

Proposition 3. *Assume Assumption 2 and Assumption 3. For $\varepsilon > 0$, $u_0 \in \mathbb{R}^{d_u}$, let $C_1(u_0) := \int_{\mathcal{X}^n} \|T(u_0,X)\|_{\infty,2}\mathrm{d}\mathbb{x}^{\otimes n}(X)$ and $C_2 := \int_{\mathcal{Y}^n} \|Y\|_{\infty,2}\mathrm{d}\mathbb{y}^{\otimes n}(Y)$.*

Let $K_F(\varepsilon,u_0) := 4Ln(\varepsilon L + C_1(u_0) + C_2)$. We have $\forall u,u' \in B(u_0,\varepsilon)$, $|F(u) - F(u')| \leq K_F(\varepsilon,u_0)\|u - u'\|_2$.

Proof. Let $\varepsilon > 0$, $u_0 \in \mathbb{R}^{d_u}$ and $u,u' \in B(u_0,\varepsilon)$. We have

$$\begin{aligned}|F(u) - F(u')| &\leq \int_{\mathcal{X}^n\times\mathcal{Y}^n\times\mathbb{S}^{d_y-1}} |f(u,X,Y,\theta) - f(u',X,Y,\theta)|\mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y)\mathrm{d}\varpi(\theta) \\ &\leq \int_{\mathcal{X}^n\times\mathcal{Y}^n} K_f(\varepsilon,u_0,X,Y)\|u-u'\|_2\mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y) \\ &\leq \int_{\mathcal{X}^n\times\mathcal{Y}^n} 4Ln(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2})\|u-u'\|_2\mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y).\end{aligned}$$

Now by Assumption 2, $X \longmapsto \|T(u_0,X)\|_{\infty,2}$ is continuous on the compact $\mathcal{X}^n$, thus upper-bounded by a certain $M(u_0) > 0$. We can define $C_1(u_0) := \int_{\mathcal{X}^n} \|T(u_0,X)\|_{\infty,2}\mathrm{d}\mathbb{x}^{\otimes n}(X)$, which verifies $C_1(u_0) \leq M(u_0)\mathbb{x}(\mathcal{X})^n$. Since $\mathcal{X}$ is compact and $\mathbb{x}$ is a Radon probability measure by Assumption 3, $\mathbb{x}(\mathcal{X})$ is well-defined and finite, thus $C_1(u_0)$ is finite. Likewise, let $C_2 := \int_{\mathcal{Y}^n} \|Y\|_{\infty,2}\mathrm{d}\mathbb{y}^{\otimes n}(Y) < +\infty$.

Finally, $|F(u) - F(u')| \leq 4Ln(\varepsilon L + C_1(u_0) + C_2)\|u - u'\|_2$. □

Having shown that our losses are locally Lipschitz, we can now turn to convergence results. These conclusions are placed in the context of non-smooth and non-convex optimisation, thus will be tied to the Clarke sub-differential of F, which we denote $\partial_C F$. The set of Clarke sub-gradients at a point u is the convex hull of the limits of gradients of F:

$$\partial_C F(u) := \mathrm{conv}\left\{v \in \mathbb{R}^{d_u} : \exists (u^{(t)}) \in (\mathcal{D}_F)^{\mathbb{N}} : u^{(t)} \xrightarrow[t\longrightarrow+\infty]{} u \text{ and } \nabla F(u^{(t)}) \xrightarrow[t\longrightarrow+\infty]{} v\right\}, \tag{8}$$

where $\mathcal{D}_F$ is the set of differentiability of F. At points u where F is differentiable, $\partial_C F(u) = \{\nabla F(u)\}$, and if F is convex in a neighbourhood of u, then the Clarke differential at u is the set of its convex sub-gradients. The interested reader may turn to Section C for further context on non-smooth and non-convex optimisation.

3 Convergence of Interpolated SGD Trajectories on F

In general, the idea behind SGD is a discretisation of the gradient flow equation $\dot{u}(s) = -\nabla F(u(s))$. In our non-smooth setting, the underlying continuous-time problem is instead the Clarke differential inclusion $\dot{u}(s) \in -\partial_C F(u(s))$. Our objective is to show that in a certain sense, the SGD trajectories approach the set of solutions of this inclusion problem, as the step size decreases. We consider solutions that are absolutely continuous (we will write $u(\cdot) \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u})$) and start within $\mathcal{K} \subset \mathbb{R}^{d_u}$, a fixed compact set. We can now define the solution set formally as

$$S_{-\partial_C F}(\mathcal{K}) := \left\{ u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u}) \mid \underline{\forall} s \in \mathbb{R}_+,\ \dot{u}(s) \in -\partial_C F(u(s));\ u(0) \in \mathcal{K} \right\}, \tag{9}$$

where we write $\underline{\forall}$ for "almost every". In order to compare the discrete SGD trajectories to this set of continuous-time trajectories, we interpolate the discrete points in an affine manner: Equation (10) defines the *piecewise-affine interpolated SGD trajectory* associated to a discrete SGD trajectory $(u_\alpha^{(t)})_{t\in\mathbb{N}}$ of learning rate α.

$$u_\alpha(s) = u_\alpha^{(t)} + \left(\frac{s}{\alpha} - t\right)(u_\alpha^{(t+1)} - u_\alpha^{(t)}), \quad \forall s \in [t\alpha, (t+1)\alpha[, \quad \forall t \in \mathbb{N}. \tag{10}$$

In order to compare our interpolated trajectories with the solutions, we consider the metric of uniform convergence on all segments

$$d_c(u, u') := \sum_{k\in\mathbb{N}^*} \frac{1}{2^k} \min\left(1, \max_{s\in[0,k]} \|u(s) - u'(s)\|_2\right). \tag{11}$$

In order to prove a convergence result on the interpolated trajectories, we will leverage the work of Bianchi et al. (2022) which hinges on three conditions on the loss F that we reproduce and verify successively. Firstly, Condition 1 assumes mild regularity on the sample loss function f.

Condition 1.

i) There exists $\kappa : \mathbb{R}^{d_u} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$ measurable such that each $\kappa(u, \cdot)$ is $\mathbb{z}$-integrable, and:

$$\exists \varepsilon > 0,\ \forall u, u' \in B(u_0, \varepsilon),\ \forall z \in \mathcal{Z},\ |f(u, z) - f(u', z)| \leq \kappa(u_0, z)\|u - u'\|_2.$$

ii) There exists $u \in \mathbb{R}^{d_u}$ such that $f(u, \cdot)$ is $\mathbb{z}$-integrable.

Our regularity result on f Proposition 2 allows us to verify Condition 1, by letting $\varepsilon := 1$ and $\kappa(u_0, z) := K_f(1, u_0, X, Y)$. Condition 1 ii) is immediate since for *all* $u \in \mathbb{R}^{d_u}$, $(X, Y, \theta) \longmapsto w_\theta(T(u, X), Y)$ is continuous in each variable separately, thanks to the regularity of T provided by Assumption 2, and to the regularities of w. This continuity implies that all $f(u, \cdot)$ are $\mathbb{z}$-integrable, since $\mathbb{z} = \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \mathbb{\sigma}$ is a compactly supported probability measure under Assumption 3. Secondly, Condition 2 concerns the local Lipschitz constant κ introduced in Condition 1: it is assumed to increase slowly with respect to the network parameters u.

Condition 2. *The function κ of Condition 1 verifies:*

i) There exists $c \geq 0$ such that $\forall u \in \mathbb{R}^{d_u}$, $\displaystyle\int_{\mathcal{Z}} \kappa(u, z)\mathrm{d}\mathbb{z}(z) \leq c(1 + \|u\|_2)$.

ii) For every compact $\mathcal{K} \subset \mathbb{R}^{d_u}$, $\displaystyle\sup_{u\in\mathcal{K}} \int_{\mathcal{Z}} \kappa(u, z)^2 \mathrm{d}\mathbb{z}(z) < +\infty$.

Condition 2.ii) is verified by κ given its regularity. However, Condition 2.i) requires that $T(u, x)$ increase slowly as $\|u\|_2$ increases, which is more costly.

Assumption 4. *There exists an $\mathbb{x}$-integrable function $g : \mathbb{R}^{d_x} \longrightarrow \mathbb{R}_+$ such that $\forall u \in \mathbb{R}^{d_u}$, $\forall x \in \mathbb{R}^{d_x}$, $\|T(u, x)\|_2 \leq g(x)(1 + \|u\|_2)$.*

Assumption 4 is satisfied in particular as soon as $T(\cdot, x)$ is bounded (which is the case for a neural network with bounded activation functions), or if T is of the form $T(u,x) = \widetilde{T}(u,x)\mathbb{1}_{B(0,R)}(u)$, i.e. limiting the network parameters u to be bounded. This second case does not yield substantial restrictions in practice (see Section D for a class of NNs that satisfy all of the assumptions), yet vastly simplifies theory. Under Assumption 4, we have for any $u \in \mathbb{R}^{d_u}$, with $\kappa(u,z) = K_f(1,u,X,Y)$ from Proposition 2 and C_2 from Proposition 3,

$$\begin{aligned}\int_{\mathcal{X}^n\times\mathcal{Y}^n\times\mathbb{S}^{d_y-1}} K_f(1,u,X,Y)\mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y)\mathrm{d}\mathbb{\sigma}(\theta) &\leq 4Ln\left(\varepsilon L + (1+\|u\|_2)\int_{\mathcal{X}^n}\max_{k\in[\![1,n]\!]} g(x_k)\mathrm{d}\mathbb{x}^{\otimes n}(X) + C_2\right)\\ &\leq c(1+\|u\|_2).\end{aligned}$$

As a consequence, Condition 2 holds under our assumptions. We now consider the Markov kernel associated to the SGD schemes:

$$P_\alpha : \left\{\begin{array}{ccc}\mathbb{R}^{d_u}\times\mathcal{B}(\mathbb{R}^{d_u}) & \longrightarrow & [0,1]\\ u, B & \longmapsto & \displaystyle\int_{\mathcal{Z}}\mathbb{1}_B(u-\alpha\varphi(u,z))\mathrm{d}\mathbb{z}(z)\end{array}\right. .$$

Given $u \in \mathbb{R}^{d_u}$, $P_\alpha(u,\cdot)$ is a probability measure on $\mathbb{R}^{d_u}$ which dictates the law of the positions of the next SGD iteration $u^{(t+1)}$, conditionally to $u^{(t)} = u$. With $\mathbb{\lambda}_{\mathbb{R}^{d_u}}$ denoting the Lebesgue measure on $\mathbb{R}^{d_u}$, let $\Gamma := \{\alpha \in]0,+\infty[\mid \forall \mathbb{u} \ll \mathbb{\lambda}_{\mathbb{R}^{d_u}},\ \mathbb{u}P_\alpha \ll \mathbb{\lambda}_{\mathbb{R}^{d_u}}\}$. Γ is the set of learning rates α for which the kernel P_α maps any absolutely continuous probability measure $\mathbb{u}$ to another such measure. We will verify the following condition, which can be interpreted as the SGD trajectories continuing to explore the entire space for a small enough learning rate α:

Condition 3. *The closure of Γ contains 0.*

In order to satisfy Condition 3, we require an additional regularity condition on the neural network T which we formulate in Assumption 5.

Assumption 5. *There exists a constant $M > 0$, such that (with the notations of Assumption 1 and Assumption 3) $\forall x \in \mathcal{X}$, $\forall j \in J(x)$, $\forall u \in \mathcal{U}_j(x)$, $\forall (i_1,i_2,i_3,i_4) \in [\![1,d_u]\!]^2 \times [\![1,d_y]\!]^2$,*

$$\left|\frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\Big([T(u,x)]_{i_3}[T(u,x)]_{i_4}\Big)\right| \leq M, \text{ and } \left\|\frac{\partial^2 T}{\partial u_{i_1}\partial u_{i_2}}(u,x)\right\|_2 \leq M.$$

The upper bounds in Assumption 5 bear strong consequences on the behaviour of T for $\|u\|_2 \gg 1$, and are only practical for networks of the form $T(u,x) = \widetilde{T}(u,x)\mathbb{1}_{B(0,R)}(u,x)$, similarly to Assumption 4. We detail the technicalities of verifying this assumption along with the others in the Appendix (Section D).

Proposition 4. *Under Assumption 1, Assumption 3 and Assumption 5, for the SGD trajectories (7), Γ contains $]0,\alpha_0[$, where $\alpha_0 := \Big((d_y{}^2 + 2R_y)d_u M\Big)^{-1}$.*

We postpone the proof to Section B. Now that we have verified Condition 1, Condition 2 and Condition 3, we can apply (Bianchi et al., 2022), Theorem 2 to F, showing a convergence result on interpolated SGD trajectories.

Theorem 1. *Consider a neural network T and measures $\mathbb{x}$, $\mathbb{y}$ satisfying Assumption 1, Assumption 2, Assumption 3, Assumption 4 and Assumption 5. Let $\alpha_1 < \alpha_0$ (see Proposition 4).*

Let $(u_\alpha^{(t)}), \alpha \in]0,\alpha_1], t \in \mathbb{N}$ a collection of SGD trajectories associated to (7). Consider (u_α) their associated interpolations. For any compact $\mathcal{K} \subset \mathbb{R}^{d_u}$ and any $\eta > 0$, we have:

$$\lim_{\substack{\alpha\longrightarrow 0\\ \alpha\in\ ]0,\alpha_1]}} \mathbb{u}_0\otimes\mathbb{x}^{\otimes\mathbb{N}}\otimes\mathbb{y}^{\otimes\mathbb{N}}\otimes\mathbb{\sigma}^{\otimes\mathbb{N}}\left(d_c(u_\alpha, S_{-\partial_C F}(\mathcal{K})) > \eta\right) = 0. \tag{12}$$

The distance d_c is defined in (11). As the learning rate decreases, the interpolated trajectories approach the trajectory set $S_{-\partial_C F}$, which is essentially a solution of the *gradient flow equation* $\dot{u}(s) = -\nabla F(u(s))$ (ignoring the set of non-differentiability, which is $\mathbb{1}_{\mathbb{R}^{d_u}}$-null). To get a tangible idea of the concepts at play, if F was $\mathcal{C}^2$ and had a finite amount of critical points, then one would have the convergence of a solution $u(s)$ to a critical point of F, as $s \longrightarrow +\infty$. These results have implicit consequences on the value of the parameters at the "end" of training for low learning rates, which is why we will consider a variant of SGD for which we can say more precise results on the convergence of the parameters.

4 Convergence of Noised Projected SGD Schemes on F

In practice, it is seldom desirable for the parameters of a neural network to reach extremely large values during training. Weight clipping is a common (although contentious) method of enforcing that $T(u, \cdot)$ stay Lipschitz, which is desirable for theoretical reasons. For instance the 1-Wasserstein duality in Wasserstein GANs (Arjovsky et al., 2017) requires Lipschitz networks, and similarly, Sliced-Wasserstein GANs (Deshpande et al., 2018) use weight clipping and enforce their networks to be Lipschitz.

Given a radius $r > 0$, we consider SGD schemes that are restricted to $u \in \overline{B}(0, r) =: B_r$, by performing *projected* SGD. At each step t, we also add a noise $a\varepsilon^{(t+1)}$, where $\varepsilon^{(t+1)}$ is an additive noise of law $\mathfrak{c} \ll \mathbb{1}_{\mathbb{R}^u}$, which is often taken as standard Gaussian in practice. These additions yield the following SGD scheme:

$$\begin{aligned} u^{(t+1)} &= \pi_r\left(u^{(t)} - \alpha\varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}) + \alpha a\varepsilon^{(t+1)}\right), \\ \left(u^{(0)}, (X^{(t)})_{t\in\mathbb{N}}\ (Y^{(t)})_{t\in\mathbb{N}},\ (\theta^{(t)})_{t\in\mathbb{N}},\ (\varepsilon^{(t)})_{t\in\mathbb{N}}\right) &\sim \mathfrak{u}_0 \otimes \mathfrak{x}^{\otimes\mathbb{N}} \otimes \mathfrak{y}^{\otimes\mathbb{N}} \otimes \mathfrak{o}^{\otimes\mathbb{N}} \otimes \mathfrak{c}^{\otimes\mathbb{N}}, \end{aligned} \tag{13}$$

where $\pi_r : \mathbb{R}^u \longrightarrow B_r$ denotes the orthogonal projection on the ball $B_r := \overline{B}(0, r)$. Thanks to Condition 1, Condition 2 and the additional noise, we can verify the assumptions for (Bianchi et al., 2022) Theorem 4, yielding the same result as Theorem 1 for the noised projected scheme (13). In fact, under additional assumptions, we shall prove a stronger mode of convergence for the aforementioned trajectories. The natural context in which to perform gradient descent is on functions that admit a chain rule, which is formalised in the case of almost-everywhere differentiability by the notion of *path differentiability*, as studied thoroughly in (Bolte & Pauwels, 2021). We also provide a brief presentation in the Appendix (Section C.1).

Condition 4. *F is path differentiable, which is to say that for any $u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u})$, for almost all $s > 0$, $\forall v \in \partial_C F(u(s))$, $v^\top \dot{u}(s) = (F \circ u)'(s)$.*

Remark 1. *There are alternate equivalent formulations for Condition 4. Indeed, as presented in further detail in Section C.1, F is path differentiable if and only if $\partial_C F$ is a conservative field for F if and only if F has a chain rule for ∂_C (the latter is the formulation chosen above in Condition 4).*

In order to satisfy Condition 4, we need to make the assumption that the NN input measure $\mathfrak{x}$ and the data measure $\mathfrak{y}$ are discrete measures, which is the case for $\mathfrak{y}$ in the case of generative neural networks, but is less realistic for $\mathfrak{x}$ in practice. We define Σ_n the n-simplex: its elements are the $a \in \mathbb{R}^n$ s.t. $\forall i \in [\![1, n]\!]$, $a_i \geq 0$ and $\sum_i a_i = 1$.

Assumption 6. *One may write $\mathfrak{x} = \sum_{k=1}^{n_x} a_k \delta_{x_k}$ and $\mathfrak{y} = \sum_{k=1}^{n_y} b_k \delta_{y_k}$, with the coefficient vectors $a \in \Sigma_{n_x}, b \in \Sigma_{n_y}$, $\mathcal{X} = \{x_1, \cdots, x_{n_x}\} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} = \{y_1, \cdots, y_{n_y}\} \subset \mathbb{R}^{d_y}$.*

There is little practical reason to consider non-uniform measures, however the generalisation to any discrete measure makes no theoretical difference. Note that Assumption 3 is clearly implied by Assumption 6.

In order to show that F is path differentiable, we require the natural assumption that each $T(\cdot, x)$ be path differentiable. Since $T(\cdot, x)$ is a vector-valued function, we need to extend the notion of path-differentiability. Thankfully, Bolte & Pauwels (2021) define *conservative mappings* for vector-valued locally Lipschitz functions (Definition 4), which allows us to define naturally path differentiability of a vector-valued function as the path-differentiability of all of its coordinate functions. See Section C.2 for a detailed presentation.

Assumption 7. *For any $x \in \mathbb{R}^{d_x}$, $T(\cdot, x)$ is path differentiable.*

Assumption 7 holds as soon as each the neural network has the typical structure of compositions of linear units and typical activations, as was proved by Davis et al. (2020), Corollary 5.11 and Bolte & Pauwels (2021), Section 6.2. We provide a more specific class of NNs that are path differentiable and satisfy all our other assumptions in Section D.

Proposition 5. *Under Assumption 2, Assumption 6 and Assumption 7, F is path differentiable.*

Proof. We shall use repeatedly the property that the composition of path differentiable functions remains path differentiable, which is proved in (Bolte & Pauwels, 2021), Lemma 6.

Let $\mathcal{E} : \begin{cases} \mathbb{R}^{n\times d_y} \times \mathbb{R}^{n\times d_y} & \longrightarrow & \mathbb{R}_+ \\ Y, Y' & \longmapsto & \mathrm{SW}_2^2(\mathfrak{b}_Y, \mathfrak{b}_{Y'}) \end{cases}$. By (Tanguy et al., 2023), Proposition 2.4.3, each $\mathcal{E}(\cdot, Y)$ is semi-concave and thus is path differentiable (by (Tanguy et al., 2023), Proposition 4.3.3).

Thanks to Assumption 6, $\mathbb{x}^{\otimes n}$ and $\mathbb{y}^{\otimes n}$ are discrete measures on $\mathbb{R}^{n\times d_x}$ and $\mathbb{R}^{n\times d_y}$ respectively, allowing one to write $\mathbb{x}^{\otimes n} = \sum_k a_k \delta_{X_k}$ and $\mathbb{y}^{\otimes n} = \sum_l b_l \delta_{Y_l}$. Then $F = u \longmapsto \sum_{k,l} a_k b_l \mathcal{E}(T(u, X_k), Y_l)$ is path differentiable as a sum ((Bolte & Pauwels, 2021), Corollary 4) of compositions ((Bolte & Pauwels, 2021), Lemma 6) of path differentiable functions. □

We have now satisfied all the assumptions to apply (Bianchi et al., 2022), Theorem 6, showing that trajectories of (13) converge towards to a set of generalised critical points[2] $\mathcal{C}_r$ defined as

$$\mathcal{C}_r := \left\{ u \in \mathbb{R}^{d_u} \mid 0 \in -\partial_C F(u) - \mathcal{N}_r(u) \right\}, \quad \mathcal{N}_r(u) = \begin{cases} \{0\} \text{ if } \|u\|_2 < r \\ \{su \mid s \geq 0\} \text{ if } \|u\|_2 = r \\ \varnothing \text{ if } \|u\|_2 > r \end{cases}, \tag{14}$$

where $\mathcal{N}_r(u)$ refers to the *normal cone* of the ball $\overline{B}(0, r)$ at x. The term $\mathcal{N}_r(u)$ in (14) only makes a difference in the pathological case $\|u\|_2 = r$, which never happens in practice since the idea behind projecting is to do so on a very large ball, in order to avoid gradient explosion, to limit the Lipschitz constant and to satisfy theoretical assumptions. Omitting the $\mathcal{N}_r(u)$ term, and denoting $\mathcal{D}$ the points where F is differentiable, (14) simplifies to $\mathcal{C}_r \cap \mathcal{D} = \{u \in \mathcal{D} \mid \nabla F(u) = 0\}$, i.e. the critical points of F for the usual differential. Like in Theorem 1, we let $\alpha_1 < \alpha_0$, where α_0 is defined in Proposition 4. We have met the conditions to apply Bianchi et al. (2022), Theorem 6, showing a long-run convergence results on the SGD trajectories (13).

Theorem 2. *Consider a neural network T and measures $\mathbb{x}$, $\mathbb{y}$ satisfying Assumption 1, Assumption 2, Assumption 4, Assumption 5, Assumption 6 and Assumption 7. Let $(u_\alpha^{(t)})_{t\in\mathbb{N}}$ be SGD trajectories defined by (13) for $r > 0$ and $\alpha \in]0, \alpha_1]$. One has*

$$\forall \eta > 0, \ \overline{\lim_{t\longrightarrow+\infty}} \ \mathbb{u}_0 \otimes \mathbb{x}^{\otimes\mathbb{N}} \otimes \mathbb{y}^{\otimes\mathbb{N}} \otimes \mathbb{\varpi}^{\otimes\mathbb{N}} \otimes \mathbb{\epsilon}^{\otimes\mathbb{N}} \left(d(u_\alpha^{(t)}, \mathcal{C}_r) > \eta \right) \xrightarrow[\substack{\alpha\longrightarrow 0 \\ \alpha\in]0,\alpha_1]}]{} 0.$$

The distance d above is the usual euclidean distance. Theorem 2 shows essentially that as the learning rate approaches 0, the long-run limits of the SGD trajectories approach the set of $\mathcal{C}_r$ in probability. Omitting the points of non-differentiability and the pathological case $\|u\|_2 = r$, the general idea is

[2] Typically referred to as the set of *Karush-Kahn-Tucker* points of the differential inclusion $\dot{u}(s) \in -\partial_C F(u(s)) - \mathcal{N}_r(u(s))$.

that $u_\alpha^{(\infty)} \xrightarrow[\alpha \longrightarrow 0]{} \{u \ : \ \nabla F(u) = 0\}$, which is the convergence that would be achieved by the gradient flow of F, in the simpler case of $\mathcal{C}^2$ smoothness.

5 Conclusion and Outlook

Under reasonable assumptions, we have shown that SGD trajectories of parameters of generative NNs with a minibatch SW loss converge towards the desired sub-gradient flow solutions, implying in a weak sense the convergence of said trajectories. Under stronger assumptions, we have shown that trajectories of a mildly modified SGD scheme converge towards a set of generalised critical points of the loss, which provides a missing convergence result for such optimisation problems.

The core limitation of this theoretical work is the assumption that the input data measure $\mathbb{x}$ is discrete (Assumption 6), which we required in order to prove that the loss F is path differentiable. In order to generalise to a non-discrete measure, one would need to apply or show a result on the stability of path differentiability through integration: in our case, we want to show that $\int_{\mathcal{X}^n} \mathcal{E}(T(u, X), Y) \mathrm{d}\mathbb{x}^{\otimes n}(X)$ is path differentiable, knowing that $u \longmapsto \mathcal{E}(T(u, X), Y)$ is path differentiable by composition (see the proof of Proposition 5 for the justification). Unfortunately, in general if each $g(\cdot, x)$ is path differentiable, it is not always the case that $\int g(\cdot, x)\mathrm{d}x$ is path differentiable (at the very least, there is no theorem stating this, even in the simpler case of another sub-class of path differentiable functions, see (Bianchi et al., 2022), Section 6.1). However, there is such a theorem (specifically (Clarke, 1990), Theorem 2.7.2 with Remark 2.3.5) for *Clarke regular* functions (see Section C.3 for a presentation of this regularity class), sadly the composition of Clarke regular functions is not always Clarke regular, it is only known to be the case in excessively restrictive cases (see (Clarke, 1990), Theorems 2.3.9 and 2.3.10). Similarly to the continuous case, the simpler generalisation in which $\mathbb{x}$ has a countable support adds substantial difficulty, since all of the typical tools (path differentiability itself, Clarke regularity or even definability (see (Bolte & Pauwels, 2021) Section 4.1 for a first introduction) do not have readily applicable results for infinite operations, to our knowledge. As a result, we leave the generalisation to a non-discrete input measure $\mathbb{x}$ for future work.

Our studies focus on the 2-SW distance, but our results from Section 3 can be extended to $p \in [1, +\infty[$, as presented in the appendix (Section E). However, as also discussed in the Appendix, the generalisation of Section 4 is still an open problem, since it has not yet be proven that $X \longmapsto \mathrm{SW}_p^p(\mathbb{X}, \mathbb{Y})$ is path differentiable for $p \neq 2$.

This paper studies the use of the *average* SW distance as a loss, and an extension to related distances would be worth considering. The average SW distance aggregates the projected distances through an expectation, while the closely-related *max*-Sliced Wasserstein distance introduced by Deshpande et al. (2019) aggregates the projections via a maximisation on the axis $\theta \in \mathbb{S}^{d-1}$. The training paradigm presented in (Deshpande et al., 2019) differs strongly from our formalism since it applies to GANs, however one could consider an extension of our formalism in which the optimal projection θ becomes a learned parameter of the neural network. A related extension is the Subspace-Robust Wasserstein distance (Paty & Cuturi, 2019), which can take the following formulation

$$\mathcal{S}_k^2(\mathbb{x}, \mathbb{y}) = \max_{\substack{0 \preceq \Omega \preceq I_d \\ \mathrm{trace}(\Omega) = k}} \mathrm{W}_2^2(\Omega^{1/2} \# \mathbb{x}, \Omega^{1/2} \# \mathbb{y}),$$

for which one could consider a similar extension where the positive semi-definite Ω becomes a learned parameter of T.

Another avenue for future study would be to tie the flow approximation result from Theorem 1 to Sliced Wasserstein Flows (Liutkus et al., 2019; Bonet et al., 2022). The difficulty in seeing the differential inclusion (9) as a flow of F lies in the non-differentiable nature of the functions at play, as well as the presence of the composition between SW and the neural network T, which bodes poorly with Clarke sub-differentials.

Acknowledgements

We thank Julie Delon for proof-reading and general feedback, as well as Rémi Flamary and Alain Durmus for fruitful discussions.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. 2021.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *arXiv*, December 2017. doi: 10.48550/arXiv.1712.01504.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30 (3):1117–1147, 2022.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188:19–51, 2021.

Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Efficient gradient flows in sliced-Wasserstein space. *Transactions on Machine Learning Research*, 2022.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Nicolas Bonnotte. Unidimensional and evolution methods for optimal transportation. *PhD Thesis, Paris 11*, 2013.

Pierre Bréchet, Katerina Papagiannouli, Jing An, and Guido Montúfar. Critical points and convergence analysis of generative deep linear networks trained with Bures-Wasserstein loss. *arXiv preprint arXiv:2303.03027*, 2023.

Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

M Coste. An introduction to o-minimal geometry, inst. rech. *RAAG Notes, Institut de Recherche Mathématique de Rennes*, 1999.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced Wasserstein distance. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3483–3491. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00367. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Deshpande_Generative_Modeling_Using_CVPR_2018_paper.html.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for

gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced Wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9412–9420, 2021.

Yu-Jui Huang, Shih-Chun Lin, Yu-Chih Huang, Kuan-Hui Lyu, Hsin-Hua Shen, and Wan-Yi Lin. On characterizing optimal Wasserstein GAN solutions for non-Gaussian data. 2023.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4104–4113. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liutkus19a.html.

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20802–20812. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/eefc9e10ebdc4a2333b42b2dbb8f27b6-Paper.pdf.

Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International conference on machine learning*, pp. 5072–5081. PMLR, 2019.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 51(1):1–44, 2019. doi: 10.1561/2200000073. URL https://arxiv.org/abs/1803.00567.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, 2012.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Eloi Tanguy, Rémi Flamary, and Julie Delon. Properties of discrete sliced Wasserstein losses. *arXiv preprint arXiv:2307.10352*, 2023.

Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016. doi: 10.1137/16M1067494. URL https://doi.org/10.1137/16M1067494.

Lou Van Den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

Cédric Villani. *Optimal transport : old and new / Cédric Villani*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

Seiichiro Wakabayashi. Remarks on semi-algebraic functions, January 2008. Online Notes.

J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. Paudel, and L. Van Gool. Sliced Wasserstein generative models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3708–3717, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00383. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00383.

Jiaqi Xi and Jonathan Niles-Weed. Distributional convergence of the sliced Wasserstein process. *Advances in Neural Information Processing Systems*, 35:13961–13973, 2022.

Xianliang Xu and Zhongyi Huang. Central limit theorem for the sliced 1-Wasserstein distance and the max-sliced 1-Wasserstein distance. *arXiv preprint arXiv:2205.14624*, 2022.

A Table of Notations

Table 1: List of Notations

Symbol	Explanation
$\mathbb{\delta}X$	Given $X = (x_1, \cdots, x_n) \in \mathbb{R}^{n\times d}$, $\mathbb{\delta}X = \frac{1}{n}\sum_i \delta_{x_i}$
X	$(x_1, \cdots, x_n) \in \mathbb{R}^{n\times d_x}$ an input data sample of law $\mathbb{x}^{\otimes n}$
$\mathbb{x}$	input data probability measure on $\mathbb{R}^{d_x}$, supported on $\mathcal{X}$
Y	$(y_1, \cdots, y_n) \in \mathbb{R}^{n\times d_y}$ a target data sample of law $\mathbb{y}^{\otimes n}$
$\mathbb{y}$	target data probability measure on $\mathbb{R}^{d_y}$, supported on $\mathcal{Y}$
θ	direction in $\mathbb{S}^{d_y-1}$
$\mathbb{\sigma}$	uniform measure on $\mathbb{S}^{d_y-1}$
$z := (X, Y, \theta)$	sample in X, Y and θ
$\mathbb{z} := \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \mathbb{\sigma}$	probability measure for the samples z, supported on $\mathcal{Z} := \mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y-1}$
u	neural network parameters in $\mathbb{R}^{d_u}$
$T(u, X)$	neural network function defined in (3)
$f(u, X, Y, \theta)$	sample loss function defined in (4)
$F(u)$	population loss function defined in (5)
$w_\theta(Y, Y')$	discrete and projected 2-Wasserstein distance $\mathrm{W}_2^2(P_\theta \# \mathbb{\delta}_Y, P_\theta \# \mathbb{\delta}_{Y'})$
$\varphi(u, X, Y, \theta)$	almost-everywhere gradient of $f(\cdot, X, Y, \theta)$ defined in (6)
K_w, K_f, K_F	local Lipschitz constants of w, f, F respectively (see Propositions 1, 2, 3)
$\alpha; a$	SGD learning rate; noise level
$\mathbb{\lambda}_{\mathbb{R}^d}; \rho \ll \mathbb{\lambda}_{\mathbb{R}^d}$	Lebesgue measure on $\mathbb{R}^d$; a measure ρ absolutely continuous w.r.t. $\mathbb{\lambda}_{\mathbb{R}^d}$
∂_C	Clarke differential, defined in (8)
$\mathbb{u}_0$	probability measure of SGD initialisation $u^{(0)}$
$\varepsilon^{(t)}$	additive noise in $\mathbb{R}^{d_u}$ at SGD step t
$\mathbb{e}$	additive noise probability measure on $\mathbb{R}^{d_u}$
$B_{\|\cdot\|}(x, R), \overline{B}_{\|\cdot\|}(x, R)$	open (resp. closed) ball of centre x and radius R for the norm $\|\cdot\|$

B Postponed Proofs

Proof of Proposition 4

Proof. Let $\mathbb{u} \ll \mathbb{\lambda}$ and $B \in \mathcal{B}(\mathbb{R}^{d_u})$ such that $\mathbb{\lambda}(B) = 0$. We have, with $\alpha' := 2\alpha/n$, $z := (X, Y, \theta)$, $\mathbb{z} := \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \mathbb{\sigma}$ and $\mathcal{Z} := \mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y-1}$,

$$\mathbb{u}P_\alpha(B) = \int_{\mathbb{R}^{d_u}\times\mathcal{Z}} \mathbb{1}_B \left[u - \alpha' \sum_{k=1}^n \left(\frac{\partial T}{\partial u}(u, x_k)\right)^\top \theta\theta^\top (T(u, x_k) - y_{\sigma_\theta^{T(u,X),Y}(k)})\right] \mathrm{d}\mathbb{u}(u)\mathrm{d}\mathbb{z}(z) \quad \leq \sum_{\tau\in\mathfrak{S}_n} \int_{\mathcal{Z}} I_\tau(z)\mathrm{d}\mathbb{z}(z),$$

where $I_\tau(z) := \int_{\mathbb{R}^{d_u}} \mathbb{1}_B(\phi_{\tau,z}(u))\,\mathrm{d}\mathbb{u}(u)$, with $\phi_{\tau,z} := u - \alpha' \underbrace{\sum_{k=1}^n \left(\frac{\partial T}{\partial u}(u, x_k)\right)^\top \theta\theta^\top (T(u, x_k) - y_{\tau(k)})}_{\psi_{\tau,z}:=}$.

Let $\tau \in \mathfrak{S}_n$ and $(X, Y, \theta) \in \mathcal{Z}$. Using Assumption 1, separate $I_\tau(z) = \sum_{j\in J} \int_{\mathcal{U}_j(X)} \mathbb{1}_B(u - \psi_{\tau,z}(u))\,\mathrm{d}\mathbb{u}(u)$, where the differentiability structure $(\mathcal{U}_j(X))_{j\in J(X)}$ is obtained using the respective differentiability structures: for each $k \in [\![1, n]\!]$, Assumption 1 yields a structure $(\mathcal{U}_{j_k}(x_k))_{j_k\in J_k(x_k)}$ of $u \longmapsto T(u, x_k)$, which depends on x_k, hence the k indices.

To be precise, define for $j = (j_1, \cdots, j_n) \in J_1(x_1) \times \cdots \times J_n(x_n)$, $\mathcal{U}_j(X) := \bigcap_{k=1}^n \mathcal{U}_{j_k}(x_k)$, and $J(X) := \{(j_1, \cdots, j_n) \in J_1(x_1) \times \cdots \times J_n(x_n) \mid \mathcal{U}_j(X) \neq \varnothing\}$. In particular, for any $k \in [\![1, n]\!]$, $T(\cdot, x_k)$ is $\mathcal{C}^2$

on $\mathcal{U}_j(X)$. Notice that the derivatives are not necessarily defined on the border $\partial\mathcal{U}_j(X)$, which is of Lebesgue measure 0 by Assumption 1, thus the values of the derivatives on the border do not change the value of the integrals (the integrals may have the value $+\infty$, depending on the behaviour of $\phi_{\tau,s}$, but we shall see that they are all finite when α is small enough).

We drop the z, τ index in the notation, and focus on the properties of ϕ and ψ as functions of u. Our first objective is to determine a constant $K > 0$, independent of u, z, τ, such that ψ is K-Lipschitz on $\mathcal{U}_j(X)$.

First, let $\chi := \left\{ \begin{array}{ccc} \mathcal{U}_j(X) & \longrightarrow & \mathbb{R}^{d_u} \\ u & \longmapsto & \left(\frac{\partial T}{\partial u}(u, x_k)\right)^\top \theta\theta^\top T(u, x_k) \end{array} \right.$. The function χ is of class $\mathcal{C}^1$, therefore we determine its Lipschitz constant by upper-bounding the $\|\cdot\|_2$-induced operator norm of its differential, denoted by $\left|\left|\left| \frac{\partial \chi}{\partial u}(u) \right|\right|\right|_2$. Notice that $\chi(u) = \frac{1}{2}\frac{\partial}{\partial u}\left(\theta^\top T(u, x_k)\right)^2$.

Now $\left|\left|\left| \frac{\partial^2}{\partial u^2}\left(\theta^\top T(u, x_k)\right)^2 \right|\right|\right|_2 \leq d_u \max_{(i_1,i_2)\in[\![1,d_u]\!]^2} \left| \frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\left(\theta^\top T(u, x_k)\right)^2 \right|$, using Assumption 5 and $|\theta_i| \leq 1$,

$$\left| \frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\left(\theta^\top T(u, x_k)\right)^2 \right| \leq \sum_{(i_3,i_4)\in[\![1,d_y]\!]^2} \left| \theta_{i_3}\theta_{i_4} \frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\Big([T(u, x_k)]_{i_3}[T(u, x_k)]_{i_4}\Big) \right| \leq {d_y}^2 M.$$

We obtain that χ is $\frac{1}{2}d_u{d_y}^2 M$-Lipschitz.

Second, let $\omega : u \in \mathcal{U}_j(X) \longmapsto \left(\frac{\partial T}{\partial u}(u, x_k)\right)^\top \theta\theta^\top y_{\tau(k)}$, also of class $\mathcal{C}^1$. We re-write $\left[\frac{\partial \omega}{\partial u}(u)\right]_{i_1,i_2} = y_{\tau(k)}^\top \theta\theta^\top \frac{\partial^2 T}{\partial u_{i_1}\partial u_{i_2}}(u, x_k)$, and conclude similarly by Assumption 5 that ω is $\|y_{\tau(k)}\|_2 d_u M$-Lipschitz.

Finally, $\psi = \sum_{k=1}^{n}(\chi_k - \omega_k)$, and is therefore $K := (\frac{1}{2}{d_y}^2 + R_y)d_u n M$-Lipschitz, with R_y from Assumption 3. We have proven that $\left|\left|\left| \frac{\partial \psi}{\partial u}(u) \right|\right|\right|_2 \leq K$ for any $u \in \mathcal{U}_j(X)$, and that K does not depend on X, Y, θ, j or u.

We now suppose that $\alpha' < \frac{1}{K}$, which is to say $\alpha < \frac{n}{2K}$. Under this condition, $\phi : \mathcal{U}_j(X) \longrightarrow \mathbb{R}^{d_u}$ is injective. Indeed, if $\phi(u_1) = \phi(u_2)$, then $\|u_1 - u_2\|_2 = \alpha'\|\psi(u_1) - \psi(u_2)\|_2 \leq \alpha' K\|u_1 - u_2\|_2$, thus $u_1 = u_2$. Furthermore, for any $u \in \mathcal{U}_j(X)$, $\frac{\partial \phi}{\partial u}(u) = \mathrm{Id}_{\mathbb{R}^{d_u}} - \alpha'\frac{\partial \psi}{\partial u}(u)$, with $\left|\left|\left| \alpha'\frac{\partial \psi}{\partial u}(u) \right|\right|\right|_2 < 1$, thus the matrix $\frac{\partial \phi}{\partial u}(u)$ is invertible (using the Neumann series method). By the global inverse function theorem, $\phi : \mathcal{U}_j(X) \longrightarrow \phi(\mathcal{U}_j(X))$ is a $\mathcal{C}^1$-diffeomorphism.

Using the change-of-variables formula, we have $\int_{\mathcal{U}_j(X)} \mathbb{1}_B(\phi(u))\mathrm{d}\mathfrak{u}(u) = \int_{\mathcal{U}_j(X)} \mathbb{1}_B(u')\mathrm{d}\phi\#\mathfrak{u}(u') = \phi\#\mathfrak{u}(B)$, we have now shown that ϕ is a $\mathcal{C}^1$-diffeomorphism, thus since $\mathfrak{u} \ll \mathbb{\lambda}$, $\phi\#\mathfrak{u} \ll \mathbb{\lambda}$. ($\alpha \ll \beta$ denoting that α is absolutely continuous with respect to β). Since $\mathbb{\lambda}(B) = 0$, it follows that the integral is 0, then by sum over j, $I_\tau(z) = 0$ and finally $\mathfrak{u}P_\alpha(B) = 0$ by integration over z and sum over τ. □

C Background on Non-Smooth and Non-Convex Analysis

This work is placed within the context of non-smooth optimisation, a field of study in part introduced by Clarke with the so-called Clarke differential, which we introduced in Equation (8) (see (Clarke, 1990) for a general reference on this object). The purpose of this appendix is to present several adjacent objects that can be useful to the application of our results, even though we do not need them in order to prove our theorems.

C.1 Conservative Fields

The Clarke differential ∂_C of a locally Lipschitz function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ (defined in Equation (8)) is an example of a *set-valued map*. Such a map is a function $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ from the subsets of $\mathbb{R}^p$ to the subsets of $\mathbb{R}^q$, for instance in the case of the Clarke differential, we have the signature $\partial_C g : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. A set-valued map D is *graph closed* if its graph $\{(u, v) \mid u \in \mathbb{R}^p,\ v \in D(u)\}$ is a closed set of $\mathbb{R}^{p+q}$. A set-valued map D is said to be a *conservative field*, when it is graph closed, has non-empty compact values and for any absolutely continuous loop $\delta \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p)$ with $\gamma(0) = \gamma(1)$, we have

$$\int_0^1 \max_{v \in D(\gamma(s))} \langle \dot{\gamma}(s), v \rangle \mathrm{d}s = 0.$$

Similarly to primitive functions in calculus, one may define a function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ using a conservative field $D : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ up to an additive constant through following expression:

$$g(u) = g(0) + \int_0^1 \max_{v \in D(\gamma(s))} \langle \dot{\gamma}(s), v \rangle \mathrm{d}s, \quad \forall \gamma \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p) \text{ such that } \gamma(0) = 1 \text{ and } \gamma(1) = u. \tag{15}$$

In this case, we say that g is a *potential function* for the field D. This notion allows us to define a new regularity class: a function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ is called *path differentiable* when there exists a conservative field of which it is a potential. A standard result in non-smooth optimisation is the following equivalence between different notions of regularity:

Proposition 6. *Bolte & Pauwels (2021), Corollary 2. Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ locally Lipschitz. We have the equivalence between the following statements:*

- *g is path differentiable*
- *$\partial_C g$ is a conservative field*
- *g has a chain rule for the Clarke differential ∂_C:*

$$\forall u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^d),\ \underset{\sim}{\forall} s > 0,\ \forall v \in \partial_C g(u(s)),\ v^\top \dot{u}(s) = (g \circ u)'(s). \tag{16}$$

This equivalence justifies the terminology used in Condition 4. The reader seeking a complete presentation of conservative field theory may refer to (Bolte & Pauwels, 2021).

C.2 Conservative Mappings

The notion of conservative fields for real-valued locally Lipschitz functions $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ can be generalised to *conservative mappings* for vector-valued locally Lipschitz functions $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$, which one may see as a generalised Jacobian matrix (see (Bolte & Pauwels, 2021), Section 3.3 for further details). A set-valued map $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{q \times p}$ is a conservative mapping for such a g if

$$\forall u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^p),\ \underset{\sim}{\forall} s > 0,\ (g \circ u)'(s) = M\dot{u}(t),\ \forall M \in J(u(s)). \tag{17}$$

In this case, we shall say that g is path differentiable. Note that if each coordinate function g_i is the potential of a conservative field D_i, then the set-valued map

$$J(u) = \left\{ \begin{pmatrix} v_1^\top \\ \vdots \\ v_q^\top \end{pmatrix} \ :\ \forall i \in [\![1, q]\!],\ v_i \in D_i(u) \right\}$$

is a conservative mapping for g (although not all conservative mappings for g can be written in this manner). As a consequence, one could interpret (simplistically) vector-valued path differentiability as coordinate-wise path differentiability.

C.3 Clarke Regularity

Another notion of regularity for locally Lipschitz functions is that of *Clarke regularity*. Let $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ and $u \in \mathbb{R}^p$, g is said to be *Clarke regular* at u if the two quantities

$$g^\circ(u;v) := \limsup_{\substack{u' \to u \\ t \searrow 0}} \frac{g(u'+tv)-g(u')}{t} \quad \text{and} \quad g'(u;v) := \lim_{t \searrow 0} \frac{g(u+tv)-g(u)}{t}$$

exist and are equal for all $v \in \mathbb{R}^p$. Note that this notion implies path differentiability by (Bolte & Pauwels, 2021), Proposition 2. Clarke regularity is the central concept of Clarke's monograph (Clarke, 1990).

C.4 Semi-Algebraic Functions

In non-smooth analysis, one of the simplest regularity cases is the class of *semi-algebraic* functions, which are essentially piecewise polynomial functions defined on polynomial pieces. To be precise, a set $\mathcal{A} \subset \mathbb{R}^d$ is *semi-algebraic* if it can be written under the form

$$\mathcal{A} = \bigcup_{i=1}^{n} \bigcap_{j=1}^{m} \left\{ u \in \mathbb{R}^d \mid P_{i,j}(u) < 0,\ Q_{i,j}(u) = 0 \right\},$$

where the $P_{i,j}$ and $Q_{i,j}$ are real multivariate polynomials. A function $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ is *semi-algebraic* if its graph $\mathcal{G} := \{(u, g(u)) \mid u \in \mathbb{R}^p\}$ is semi-algebraic.

A locally Lipschitz real-valued semi-algebraic function is path differentiable (see for instance (Bolte & Pauwels, 2021), Proposition 2), and in the light of (Bolte & Pauwels, 2021), Lemma 3, this is also the case in the vector-valued case. Another useful property of semi-algebraic functions is that their class is stable by composition and product. The interested reader may consult (Wakabayashi, 2008) for additional properties of semi-algebraic objects, or (Coste, 1999; Van Den Dries & Miller, 1996), for a presentation of o-minimal structures, a generalisation of this concept.

D Suitable Neural Networks

In this section, we detail our claim that typical NN structures satisfy our conditions. To this end, we define a class of practical neural networks whose properties are sufficient (not all NNs that satisfy our assumptions are within this framework). Consider $\mathcal{T}$ the set of NNs T of the form

$$T : \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ (u,x) & \longmapsto & \widetilde{T}(u,x) \mathbb{1}^{\varepsilon}_{B(0,R_u)}(u) \mathbb{1}^{\varepsilon}_{B(0,R_x)}(x) \end{array} \right. ,$$

with $R_u, R_x > 0$ and $\varepsilon > 0$. The function $\mathbb{1}^{\varepsilon}_{B(0,R)}$ is a smoothed version of the usual indicator function $\mathbb{1}_{B(0,R)}$: it is any function that has value 1 in $B(0, R-\varepsilon)$, 0 outside $B(0, R+\varepsilon)$ and is $\mathcal{C}^2$-smooth (see Remark 2 for a possible construction). Given that one may take arbitrarily large radii, these indicators are added for theoretical purposes and impose no realistic constraints in practice. Additionally, $\widetilde{T} = h_N$, the N-th layer of a recursive NN structure defined by

$$h_0(u,x) = x, \quad \forall n \in [\![1, N]\!],\ h_n = \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_n} \\ (u,x) & \longmapsto & a_n \left(\sum_{i=0}^{n-1} A_{n,i}(u) h_i(u,x) + B_n u \right) \end{array} \right. ,$$

where:

- *All* functions $a_n : \mathbb{R} \longrightarrow \mathbb{R}$ are $\mathcal{C}^2$-smooth, or *all* locally Lipschitz semi-algebraic activation functions (applied entry-wise). The former condition is satisfied by the common sigmoid, hyperbolic tangent or softplus activations. The latter condition applies to the non-differentiable ReLU activation, its "Leaky ReLU" extension, and continuous piecewise polynomial activations. Note that other non-linearities such as softmax can also be considered under the same regularity restrictions, but we limit ourselves to entry-wise non-linearities for notational consistency.
- Each dimension d_n is a positive integer, with obviously $d_N = d_y$, the output dimension.
- Each $A_{n,i}$ is a linear map: $\mathbb{R}^{d_u} \longrightarrow \mathbb{R}^{d_n \times d_i}$, which maps a parameter vector u to a $d_n \times d_i$ matrix. Since the entire parameter vector u is given at each layer, this allows the architecture to only use certain parameters at each layer (as is more typical in practice). One may see this map as a 3-tensor of shape (d_n, d_i, d_u), as specified in the formulation

$$\forall u \in \mathbb{R}^{d_u},\ \forall h \in \mathbb{R}^{d_i},\ A_{n,i}(u)h = \left(\sum_{j_2=1}^{d_i} \sum_{j_3=1}^{d_u} A^{(n,i)}_{j_1,j_2,j_3} h_{j_2} u_{j_3} \right)_{j_1 \in [\![1,d_n]\!]} \in \mathbb{R}^{d_n}. \tag{18}$$

- The matrix $B_n \in \mathbb{R}^{d_n \times d_u}$ determines the intercept from the full parameter vector u.

In this model, each layer depends on all the previous layers, allowing for residual inputs for instance. Overall, all typical networks fit this description, once bounded using the indicator functions, with only a technicality on the regularities of the activations which need to be *all* $\mathcal{C}^2$-smooth, or *all* semi-algebraic. One could extend this class of NNs to those with *definable* activations within the same o-minimal structure (similarly to Davis et al. (2020) and Bolte & Pauwels (2021)).

Remark 2. *We mention that we may construct a $\mathcal{C}^\infty$-smooth $\mathbb{1}^\varepsilon_{B(0,R)}$ in $\mathbb{R}^d$ explicitly as follows:*

$$f(s) := \begin{cases} e^{-1/s} \text{ if } s > 0 \\ \text{else } 0 \end{cases}, \quad g(s) := \frac{f(s)}{f(s) + f(1-s)}, \quad \mathbb{1}^\varepsilon_{B(0,R)} := \begin{cases} \mathbb{R}^d & \longrightarrow & [0,1] \\ u & \longmapsto & g\left(\dfrac{(R+\varepsilon)^2 - \|u\|_2^2}{4R\varepsilon} \right) \end{cases}.$$

Before proving the properties of NNs from the class $\mathcal{T}$, we require a technical result on path differentiable functions.

Proposition 7. *Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ path differentiable, and $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ of class $\mathcal{C}^1$. Then their product fg is path differentiable.*

Proof. Our objective is to apply (Bolte & Pauwels, 2021) Corollary 2 (stated in Proposition 6), which is to say that $h := fg$ admits a chain rule for $\partial_C h$). First, we apply the definition of the Clarke differential and compute

$$\forall u \in \mathbb{R}^d,\ \partial_C f(u) = f(u)\nabla g(u) + g(u)\partial_C f(u) := \{ f(u)\nabla g(u) + g(u)v \mid v \in \partial_C f(u) \}.$$

Note that we used the smoothness of g. We now consider an absolutely continuous curve $u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^d)$. By Bolte & Pauwels (2021) Lemma 2, since f is path differentiable, $f \circ u$ is differentiable almost everywhere. Let D the associated set of differentiability, then let $s \in D$ and $v \in \partial_C h(u(s))$, writing $v = f(u(s))\nabla g(u(s)) + g(u(s))w$ with $w \in \partial_C f(u(s))$. We compute $(h \circ u)'(s) = (f \circ u)'(s) g(u(s)) + f(u(s))(g \circ u)'(s)$. Now since f is path differentiable and $w \in \partial_C f(u(s))$, by Proposition 6 item 3, we have $(f \circ u)'(s) = \langle w, \dot{u}(s) \rangle$. On the other hand, $(g \circ u)'(s) = \langle \nabla g(u(s)), \dot{u}(s) \rangle$ since g is $\mathcal{C}^1$. Finally by definition of v and bilinearity of $\langle \cdot, \cdot \rangle$,

$$(h \circ u)'(s) = \langle w, \dot{u}(s) \rangle g(u(s)) + f(u(s)) \langle \nabla g(u(s)), \dot{u}(s) \rangle = \langle v, \dot{u}(s) \rangle.$$

□

We now have all the tools to prove that the class of NNs $\mathcal{T}$ satisfies all of the assumptions of our paper.

Proposition 8. *All networks of the class $\mathcal{T}$ verify Assumption 1, Assumption 2, Assumption 4, Assumption 5 and Assumption 7.*

Proof. Let $T \in \mathcal{T}$, and $\widetilde{T}$ its associated underlying network. We begin with regularity considerations.

Verifying Assumptions 1 and 7 in the $\mathcal{C}^2$ Case In the case where the activations are $\mathcal{C}^2$-smooth, then each $\widetilde{T}(\cdot, x)$ is also of class $\mathcal{C}^2$. Furthermore, the smooth indicator $\mathbb{1}^{\varepsilon}_{B(0,R_u)}$ is $\mathcal{C}^\infty$-smooth, thus we can conclude that $T(\cdot, x)$ is $\mathcal{C}^2$-smooth, and thus satisfies Assumption 1 trivially. In particular, $T(\cdot, x)$ is path differentiable for any $x \in \mathbb{R}^{d_x}$, thus T also satisfies Assumption 7.

Verifying Assumptions 1 and 7 in the Semi-Algebraic Case In the case where the activations are locally Lipschitz and semi-algebraic, it follows that each $\widetilde{T}(\cdot, x)$ is semi-algebraic, which yields naturally a differentiability structure associated to the polynomial pieces, satisfying Assumption 1. Furthermore, this regularity yields path differentiability by (Bolte & Pauwels, 2021), Proposition 2. By product with the smooth indicator function, T is path differentiable by Proposition 7, therefore it satisfies Assumption 7 .

Verifying Assumption 2 in the $\mathcal{C}^2$ Case In the case where the activations are $\mathcal{C}^2$-smooth, it is clear that by composition and product $(u, x) \longmapsto \widetilde{T}(u, x)$ is *jointly* $\mathcal{C}^2$-smooth. As a consequence, it is Lipschitz jointly in (u, x) on any compact of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_y}$, and by product with the smooth indicators, so is T. Since T is zero outside $\overline{B}(0, R_u + \varepsilon) \times \overline{B}(0, R_x + \varepsilon)$, we conclude that it is globally Lipschitz in (u, x).

Verifying Assumption 2 in the Semi-Algebraic Case In the case of locally Lipschitz and semi-algebraic activations, we prove that $\widetilde{T}$ is jointly Lipschitz on any compact $\mathcal{K}$ by strong induction on $n \in [\![1, N]\!]$. Let $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$ a product compact of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_y}$, and P_n : "$\exists L_n > 0 : h_n$ is L_n-Lipschitz on $\mathcal{K}$". The initialisation P_0 is trivial, since $z(u, x) = x$. Let $n \in [\![1, N]\!]$ and assume P_i to hold true for $i \in [\![0, n-1]\!]$. In particular, the h_i are jointly continuous in (u, x), allowing the definition of

$$M := \max_{(u,x)\in\mathcal{K}} \left| \sum_{i=0}^{n-1} A_{n,i}(u) h_i(u, x) + B_n u \right|.$$

Since a_n is locally Lipschitz, a covering argument shows that there exists $L_{a_n} > 0$ such that a_n is L_{a_n}-Lipschitz on $[-M, M]$. Now let $(u_1, u_2) \in \mathcal{K}_1^2$ and $(x_1, x_2) \in \mathcal{K}_2^2$. We have

$$\begin{aligned} \|h_n(u_1, x_1) - h_n(u_2, x_2)\|_2 &\leq L_{a_n} \left\| \sum_{i=0}^{n-1} A_{n,i}(u_1) h_i(u_1, x_1) + B_n u_1 - \sum_{i=0}^{n-1} A_{n,i}(u_2) h_i(u_2, x_2) - B_n u_2 \right\|_2 \\ &\leq L_{a_n} \left(\|B_n\|_{\mathrm{op}} \|u_1 - u_2\|_2 + \sum_{i=0}^{n-1} \|A_{n,i}(u_1) h_i(u_1, x_1) - A_{n,i}(u_2) h_i(u_2, x_2)\|_2 \right), \end{aligned} \tag{19}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the $\|\cdot\|_2$-induced operator norm. Let $i \in [\![0, n-1]\!]$, we separate the norm:

$$\begin{aligned} \|A_{n,i}(u_1) h_i(u_1, x_1) - A_{n,i}(u_2) h_i(u_2, x_2)\|_2 \leq \|A_{n,i}(u_1) h_i(u_1, x_1) - A_{n,i}(u_2) h_i(u_1, x_1)\|_2 &=: \Delta_1 \\ + \|A_{n,i}(u_2) h_i(u_1, x_1) - A_{n,i}(u_2) h_i(u_2, x_2)\|_2 &=: \Delta_2. \end{aligned} \tag{20}$$

For Δ_1, use the tensor form (18) and the inequality $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$ for $x \in \mathbb{R}^d$, then $\|u\|_\infty \leq \|u\|_2$:

$$\begin{aligned}\Delta_1 &\leq \sqrt{d_n}\left\|\left(\sum_{j_2=1}^{d_i}\sum_{j_3=1}^{d_u} A^{(n,i)}_{j_1,j_2,j_3} h_i(u_1,x_1)_{j_2}(u_{j_3}^{(1)} - u_{j_3}^{(2)})\right)_{j_1 \in [\![1,d_n]\!]}\right\|_\infty \\ &\leq \sqrt{d_n}\max_{j_1,j_2,j_3} |A^{(n,i)}_{j_1,j_2,j_3}| \max_{(u,x)\in\mathcal{K}_1\times\mathcal{K}_2} \|h_i(u,x)\|_\infty \, \|u_1 - u_2\|_\infty \\ &\leq L_{\Delta_1}\|u_1 - u_2\|_2. \end{aligned} \tag{21}$$

For Δ_2, we leverage P_i and obtain

$$\Delta_2 \leq \max_{u\in\mathcal{K}_1} \|A_i(u)\|_{\mathrm{op}} \, \|h_i(u_1,x_1) - h_i(u_2,x_2)\|_2 \leq \max_{u\in\mathcal{K}_1} \|A_i(u)\|_{\mathrm{op}} \, L_i\left(\|u_1 - u_2\|_2 + \|x_1 - x_2\|_2\right). \tag{22}$$

Combining (19) (20) (21) and (22) shows P_n and concludes the induction, which in turn shows that $\widetilde{T}$ is jointly Lipschitz on any compact. Like in the smooth case, we conclude that T is globally Lipschitz, and thus that Assumption 2 holds.

Verifying Assumption 4 Under both cases of regularity for the activations,

$$g := x \longmapsto \max_{u\in\overline{B}(0,R_u+\varepsilon)} \|\widetilde{T}(u,x)\|_2 \mathbb{1}^{\varepsilon}_{B(0,R_x)}(x)$$

is measurable and bounded. Furthermore, observe that for $u, x \in \mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$, $\|T(u,x)\|_2 \leq g(x)$. As a consequence, Assumption 4 holds.

Verifying Assumption 5 in the $\mathcal{C}^2$ case If all activations are $\mathcal{C}^2$-smooth, both $\widetilde{T}$ and its coordinate-wise products $T_i \times T_j$ are $\mathcal{C}^2$-smooth jointly in (u,x). As a result, one may bound these terms on $(u,x) \in \overline{B}(0,R_u+\varepsilon)\times\overline{B}(0,R_x+\varepsilon)$ by a constant M, independent of u, x, and the assumption is verified.

Verifying Assumption 5 in the semi-algebraic case In the semi-algebraic case, there exists a structure $(\mathcal{U}_j)_{j\in J}$ of open sets of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$ whose closures cover the entire space, such that $\widetilde{T}$ is polynomial in (u,x) on each $\mathcal{U}_j$, with J finite (this is possible since $\widetilde{T}$ is jointly semi-algebraic). The NN can be written $T(u,x) = \widetilde{T}(u,x)\mathbb{1}^{\varepsilon}_{B(0,R_u)}(u)\mathbb{1}^{\varepsilon}_{B(0,R_x)}(x)$, and is therefore $\mathcal{C}^2$-smooth on each $\mathcal{U}_j$. Furthermore, its restriction to $\mathcal{U}_j$ is extendable $\mathcal{C}^2$-smoothly to $\overline{\mathcal{U}_j}$ (we shall not introduce a different notation to these extensions, for legibility). As a result, one may introduce the following bounds on the derivatives of the coordinate functions on the intersection of the compact $\mathcal{K} := \overline{B}(0,R_u+\varepsilon)\times\overline{B}(0,R_x+\varepsilon)$ and $\overline{\mathcal{U}_j}$: there exists an $M_j > 0$ such that

$$\forall (u,x) \in \mathcal{K} \cap \overline{\mathcal{U}_j}, \left|\frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\Big([T(u,x)]_{i_3}[T(u,x)]_{i_4}\Big)\right| \leq M_j \text{ and } \left\|\frac{\partial^2 T}{\partial u_{i_1}\partial u_{i_2}}(u,x)\right\|_2 \leq M_j.$$

Since J is finite and the $(\mathcal{U}_j)_{j\in J}$ cover $\mathcal{K}$, we deduce that this bound holds for $(u,x) \in \mathcal{K}$ for a common constant $M > 0$. Moreover, since T is the zero function outside of $\mathcal{K}$, this bounds also holds for any $(u,x) \in \mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$. Finally, this shows that Assumption 5 holds. □

E Generalisation to Other Sliced Wasserstein Orders

In this section, we shall discuss how some of our results can be extended by replacing the 2-SW term SW_2^2 with SW_p^p for $p \in [1, +\infty[$.

Determining Lipschitz Constants The first difficulty lies in showing that the functions $w_\theta^{(p)} := (X,Y) \longmapsto \mathrm{W}_p^p(P_\theta \# \mathbb{o}_X, P_\theta \# \mathbb{o}_Y)$ still have a locally Lipschitz regularity similar to Proposition 1 (this proposition is only shown for $p=2$ in (Tanguy et al., 2023)). We generalise their result in the following proposition.

Proposition 9. *Let $K_w^{(p)}(r,X,Y) := 2pn(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}$, for $X,Y \in \mathbb{R}^{n\times d_y}$ and $r>0$. Then $w_\theta^{(p)}(\cdot,Y)$ is $K_w^{(p)}(r,X,Y)$-Lipschitz in the neighbourhood $B_{\|\cdot\|_{\infty,2}}(X,r)$:*

$$\forall Y', Y'' \in B_{\|\cdot\|_{\infty,2}}(X,r),\ \forall \theta \in \mathbb{S}^{d_y-1},\ |w_\theta(Y',Y) - w_\theta(Y'',Y)| \leq K_w^{(p)}(r,X,Y)\|Y'-Y''\|_{\infty,2}.$$

Proof. Let $X,Y \in \mathbb{R}^{n\times d_y}, r>0$ and $Y',Y'' \in B_{\|\cdot\|_{\infty,2}}(X,r)$. By (Tanguy et al., 2023) Lemma 2.2.1, we have $|w_\theta^{(p)}(Y') - w_\theta^{(p)}(Y'')| \leq 2\|C'-C''\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm, and C' is a $n\times n$ matrix of entries $C'_{k,l} = |\theta^\top y'_k - \theta^\top y_l|^p$, with similarly $C''_{k,l} = |\theta^\top y''_k - \theta^\top y_l|^p$. Now consider the function

$$g_{y_l} := \left\{ \begin{array}{ccc} \mathbb{R}^{d_y} & \longrightarrow & \mathbb{R} \\ y & \longmapsto & |\theta^\top y - \theta^\top y_l|^p \end{array} \right. ,$$

which satisfies $C'_{k,l} = g_{y_l}(y'_k)$, and is differentiable almost-everywhere, with $\nabla g_{y_l}(y) = p|\theta^\top y - \theta^\top y_l|^{p-1}\theta$. For almost every $y \in B(x_k,r)$, we have

$$\|\nabla g_{y_l}(y)\|_2 \leq p\|y-y_l\|_2^{p-1} = p\|y-x_k+x_k-y_l\|_2^{p-1} \leq p\left(\|y-x_k\|_2 + \|x_k\|_2 + \|y_l\|_2\right)^{p-1} \leq p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}$$

As a result, g_{y_l} is $p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}$-Lipschitz in $B(x_k,r)$. Now since $Y', y'' \in B_{\|\cdot\|_{\infty,2}}(X,r)$, we have $y'_k, y''_k \in B(x_k,r)$, thus

$$|[C']_{k,l} - [C'']_{k,l}| = |g_{y_l}(y'_k) - g_{y_l}(y''_k)| \leq p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}\|y'_k - y''_k\|_2.$$

Then $\|C'-C''\|_F = \sqrt{\sum_{k,l}|[C']_{k,l}-[C'']_{k,l}|^2} \leq np(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}\|Y'-Y''\|_{\infty,2}$. □

Our results regarding the local Lipschitz property of f and F adapt immediately using the same method with the different constant $K_w^{(p)}(r,X,Y)$, we obtain the following constant for f (with L from Assumption 2):

$$K_f^{(p)}(\varepsilon,u_0,X,Y) = 2pnL\left(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2}\right)^{p-1},$$

then the following constant for F:

$$K_F^{(p)}(\varepsilon,u_0) = 2pnL\int_{\mathcal{X}^n\times\mathcal{Y}^n}\left(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2}\right)^{p-1} \mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y).$$

In order to satisfy Condition 2 item i) in the case $p \neq 2$, one needs to modify Assumption 4 to require $\|T(u,x)\|_2 \leq g(x)^{1/(p-1)}(1+\|u\|_2)^{1/(p-1)}$, which in realistic cases is not much more expensive than asking for T to be bounded, which is a property of the class of NNs that we present in Section D.

Almost-Everywhere Gradient A second difficulty lies in defining an almost-everywhere gradient f, since in our main text we rely on the formulation of an almost-everywhere gradient of $w_\theta^{(2)}(\cdot,Y)$ which was derived only for $p=2$ by Bonneel et al. (2015) and Tanguy et al. (2023). In fact, for θ, Y fixed $w_\theta^{(p)}(X,Y)$ is piecewise smooth, like $w_\theta^{(2)}(\cdot,Y)$ is piecewise quadratic. As a result, one may show that the following is an almost-everywhere gradient of $w_\theta^{(p)}(\cdot,Y)$:

$$\frac{\partial w_\theta^{(p)}}{\partial X}(X,Y) = \left(\frac{p}{n}\mathrm{sign}\left(\theta^\top x_k - \theta^\top y_{\sigma_\theta^{X,Y}(k)}\right)\left|\theta^\top x_k - \theta^\top y_{\sigma_\theta^{X,Y}(k)}\right|^{p-1}\theta\right)_{k\in[\![1,n]\!]} \in \mathbb{R}^{n\times d_y}.$$

The chain rule now yields the following almost-everywhere gradient for f:

$$\varphi(u,X,Y,\theta) = \sum_{k=1}^n \frac{p}{n}\mathrm{sign}\left(\theta^\top T(u,x_k) - \theta^\top y_{\sigma_\theta^{T(u,X),Y}(k)}\right)\left|\theta^\top T(u,x_k) - \theta^\top y_{\sigma_\theta^{T(u,X),Y}(k)}\right|^{p-1}\frac{\partial T}{\partial u}(u,x_k)\theta.$$

Adapting Proposition 4 Moving on to adapting Proposition 4, the general case $p \neq 2$ makes things substantially more technical, but one may still show that the ψ functions are Lipschitz using restrictions on T its first and second-order derivatives (which can be formulated in a more technical version of Assumption 5). In conclusion, Proposition 4 can be adapted to apply to $p \in [1, +\infty[$, and it follows that Theorem 1 also generalises to this case.

Path Differentiability Regarding the results from Section 4, the only substantial difference lies in showing that $T(\cdot, x)$ is path differentiable. The only missing link in the composition chain is the path differentiability of $\mathcal{E}^{(p)} := X \longmapsto \int_{\mathbb{S}^{d-1}} w_\theta^{(p)}(X, Y) \mathrm{d}\varpi(\theta)$. In the case $p = 2$, the difficulty of the integral can be circumvented by noticing that $\mathcal{E}$ is semi-concave (Tanguy et al., 2023), Proposition 2.4.3, which implies path differentiability. This argument does not generalise to $p \in [1, +\infty[$ naturally, hence our Theorem 2 only generalises to $p \in [1, +\infty[$ under the conjecture that $\mathcal{E}^{(p)}$ is indeed path differentiable.