Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses

Anonymous authors
Paper under double-blind review

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.

1 Introduction

1.1 The Sliced Wasserstein Distance 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

$$\forall \nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{W}_2^2(\nu_1, \nu_2) := \inf_{\pi \in \Pi(\nu_1, \nu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_1 - x_2\|^2 \mathrm{d}\pi(x_1, x_2), \tag{W2}$$

where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ admitting a second-order moment and where $\Pi(\nu_1, \nu_2)$ is the set of measures of $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ of first marginal ν_1 and second marginal ν_2. 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 μ_u, parametrised by 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), 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.

A less cost-intensive alternative to W_2^2 is the Sliced Wasserstein (SW) Distance introduced by Bonneel et al. (2015), which consists in computing the 1D Wasserstein distances between projections of input measures, and averaging over the projections. The aforementioned projection of a measure ν on $\mathbb{R}^d$ is done by the *push-forward* operation by the map $P_\theta : x \longmapsto \theta \cdot x$. Formally, $P_\theta\#\nu$ is the measure on $\mathbb{R}$ such that for any Borel set $B \subset \mathbb{R}$, $P_\theta\#\nu(B) = \nu(P_\theta^{-1}(B))$. Once the measures are projected onto a line $\mathbb{R}\theta$, the computation of the Wasserstein distance becomes substantially simpler numerically. We shall illustrate this fact in the discrete case, which arises in practical optimisation settings. Let two discrete measures on $\mathbb{R}^d$: $\gamma_X := \frac{1}{n}\sum_k \delta_{x_k}$, $\gamma_Y := \frac{1}{n}\sum_k \delta_{y_k}$ with $x_1, \cdots, x_n, y_1, \cdots, y_n \in \mathbb{R}^d$. Their push-forwards by P_θ are simply computed by the formula $P_\theta \# \gamma_X = \sum_k \delta_{P_\theta(x_k)}$, and the 2-Wasserstein distance between their projections can be computed by sorting their supports: let σ a permutation sorting $(\theta^T x_1, \cdots, \theta^T x_n)$, and τ a permutation sorting $(\theta^T y_1, \cdots, \theta^T y_n)$, one has the simple expression

$$\mathrm{W}_2^2(P_\theta \# \gamma_X, P_\theta \# \gamma_Y) = \frac{1}{n}\sum_{k=1}^{n}(\theta^T x_{\sigma(k)} - \theta^T y_{\tau(k)})^2. \tag{1}$$

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

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

The implicit generative modelling framework is a formalisation of the training step of generative Neural Networks (NNs), where a network T of parameters u is learned such as to minimise the discrepancy between $T_u \# \mathbb{x}$ [1] and $\mathbb{y}$, where $\mathbb{x}$ is a low-dimensional input distribution (often chosen as Gaussian or uniform noise), and where $\mathbb{y}$ is the target distribution. Our case of interest is when the discrepancy is measured with the SW distance, which leads to minimising $\mathrm{SW}_2^2(T_u \# \mathbb{x}, \mathbb{y})$ in u. In order to train a NN in this manner, at each iteration one draws n samples from $\mathbb{x}$ and $\mathbb{y}$ (denoted γ_X and γ_Y as discrete measures with n points), as well as a projection θ (or a batch of p projections) and performs an SGD step on the sample loss

$$\mathcal{L}(u) = \mathrm{SW}_2^2(P_\theta \# T_u \# \gamma_X, P_\theta \# \gamma_Y) = \frac{1}{n}\sum_{k=1}^{n}(\theta^T T_u(x_{\sigma(k)}) - \theta^T y_{\tau(k)})^2, \tag{2}$$

with respect to the parameters u (see Algorithm 1 for a precise formalisation). In order to compute this numerically, the main complexity comes from determining the permutations σ and τ by sorting the numbers $(\theta^T T_u(x_k))_k$ and $(y_k)_k$, and summing the results, while the Wasserstein alternative $\mathrm{W}_2^2(T_u \# \gamma_X, \gamma_Y)$ is done by solving a Linear Program, which is substantially costlier.

In this paper, we shall study this training method theoretically and prove convergence results. Theoretical guarantees for this optimisation problem are welcome, since this question has not yet been tackled (to our knowledge), even though its use is relatively widespread: for instance, Deshpande et al. (2018) and Wu et al. (2019) train GANs and auto-encoders with this method. 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).

In practice, it has been observed that SGD in such settings always converges (in the loose numerical sense), yet this property is not known theoretically, since the loss function defined in (2) is not differentiable nor convex in general, because $X \longmapsto \mathrm{SW}_2^2(\gamma_X, \gamma_Y)$ and the neural network do not have such regularities. Several efforts have been made to prove the convergence of SGD trajectories within this theoretically difficult setting: Bianchi et al. (2022) show the convergence of fixed-step SGD schemes on a function F with some technical regularity assumptions, Majewski et al. (2018) show the convergence of diminishing-step SGD schemes assuming stronger regularity results on F. Another notable theoretical work is by Bolte & Pauwels (2021), which leverages conservative field theory to prove convergence for back-propagated SGD on deep NNs with typical activations and loss functions. In the case of Optimal Transport losses, the only work (that we are aware of) that has tackled this problem is by Fatras et al. (2021), proving strong convergence results for minibatch variants of classical OT distances, namely the Wasserstein, Entropic Wasserstein and Gromov Wasserstein distances. The aim of this work is to bridge the gap between theory and practical observation by proving convergence results for SGD on Sliced Wasserstein generative losses of the form $F(u) = \mathrm{SW}_2^2(T_u \# \mathbb{x}, \mathbb{y})$.

---

[1] $T_u \# \mathbb{x}$ is the push-forward measure of $\mathbb{x}$ by T_u, i.e. the law of $T_u(x)$ when $x \sim \mathbb{x}$.

1.2 Contributions

Convergence of Interpolated SGD Under Practical Assumptions Under practically realistic assumptions, we prove in Theorem 1 that piece-wise affine interpolations (defined in Equation (7)) of constant-step SGD schemes on $u \longmapsto F(u)$ (formalised in Equation (4)) converge towards the set of sub-gradient flow solutions (see Equation (6)) 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 on any compact. We require additional assumptions on T which are more costly, but are verified as long as T is of the form $T(u, x) = \widetilde{T}(u, x)\mathbb{1}_B(u)$, where $\widetilde{T}$ is any typical NN composed of compositions of path differentiable activations (as is the case for all typical activations, see (Bolte & Pauwels, 2021), §6.2), and of linear units; and where $\mathbb{1}_B(u)$ is the indicator that the parameter u be in a fixed ball B. This form for T is a strong theoretical assumption, but in practice makes little difference, as one may take a fixed ball B to be arbitrarily large.

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 such as those of Nadjahi et al. (2020) 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

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

by minimising $u \longmapsto \mathrm{SW}_2^2(T_u\#\mathbb{x}, \mathbb{y})$ 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, $\mathbb{\sigma}$ 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 $\gamma_X := \frac{1}{n}\sum_i \delta_{x_i}$. By abuse of notation, we write $T_u(X) := (T_u(x_1), \cdots, T_u(x_n)) \in \mathbb{R}^{n \times d_y}$.

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 $f : \mathbb{R}^{d_u} \times \Xi \longrightarrow \mathbb{R}$ that is locally Lipschitz in the first variable, and ζ a probability measure on $\Xi \subset \mathbb{R}^d$ which is the law of the samples drawn at each SGD iteration. Consider $\varphi : \mathbb{R}^{d_u} \times \Xi \longrightarrow \mathbb{R}^{d_u}$ an *almost-everywhere gradient* of f, which is to say that for almost every $(u, S) \in \mathbb{R}^{d_u} \times \Xi$, $\varphi(u, S) = \partial_u f(u, S)$ (since each $f(\cdot, S)$ is locally

```
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 $\theta^{(t+1)} \sim \sigma,\ X^{(t+1)} \sim \mathbb{x}^{\otimes n}\ Y^{(t+1)} \sim \mathbb{y}^{\otimes n}$. SGD update:
    $u^{(t+1)} = u^{(t)} - \alpha \left[ \frac{\partial}{\partial u} \mathrm{W}_2^2(P_{\theta^{(t+1)}} \# T_u \# \gamma_{X^{(t+1)}}, P_{\theta^{(t+1)}} \# \gamma_{Y^{(t+1)}}) \right]_{u = u^{(t)}}$
end
```

Lipschitz, it is differentiable almost-everywhere by Rademacher's theorem). The complete loss function is $F := u \longrightarrow \int_\Xi f(u, S) \mathrm{d}\zeta(S)$. An SGD trajectory of step $\alpha > 0$ for F is a sequence $(u^{(t)}) \in (\mathbb{R}^{d_u})^{\mathbb{N}}$ of the form:

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

where ν is the distribution of the initial position $u^{(0)}$. Within this framework, we define an SGD scheme described by Algorithm 1, with $\zeta := \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \sigma$ and

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

With this definition for f, we have $F(u) = \mathbb{E}_{(X,Y,\theta) \sim \zeta} \left[\mathrm{W}_2^2(P_\theta \# T_u \# \gamma_X, P_\theta \# \gamma_Y) \right] = \mathrm{SW}_2^2(T_u \# \mathbb{x}, \mathbb{y})$: the complete loss compares the data $\mathbb{y}$ with the model's generation $T_u \# \mathbb{x}$ using 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 $Z, Y \in \mathbb{R}^{n \times d_y}$ and $\theta \in \mathbb{S}^{d_y - 1}$, $w_\theta(Y, Z) := \mathrm{W}_2^2(P_\theta \# \gamma_Z, P_\theta \# \gamma_Y)$. 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 $w_\theta(\cdot, Y)$:

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

where the permutation $\sigma_\theta^{Z,Y} \in \mathfrak{S}_n$ is $\tau_Y^\theta \circ (\tau_Z^\theta)^{-1}$, with $\tau_Y^\theta \in \mathfrak{S}_n$ being a sorting permutation of the list $(\theta \cdot y_1, \cdots, \theta \cdot y_n)$. 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 $\lambda_{\mathbb{R}^{d_u}}$ 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 $\forall j \in J(x),\ T(\cdot, x) \in \mathcal{C}^2(\mathcal{U}_j(x), \mathbb{R}^{d_y})$, $\bigcup\limits_{j \in J(x)} \overline{\mathcal{U}_j(x)} = \mathbb{R}^{d_u}$ and $\lambda_{\mathbb{R}^{d_u}}\Big(\bigcup\limits_{j \in J(x)} \partial \mathcal{U}_j(x) \Big) = 0$.*

Note that for measure-theoretic reasons, the sets $J(x)$ are assumed countable.

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 (3).

$$\varphi := \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{n \times d_x} \times \mathbb{R}^{n \times d_y} \times \mathbb{S}^{d_y - 1} & \longrightarrow & \mathbb{R}^{d_u} \\ (u, X, Y, \theta) & \longmapsto & \displaystyle\sum_{k=1}^{n} \frac{2}{n} \left(\frac{\partial T}{\partial u}(u, x_k) \right)^T \theta \theta^T (T(u, x_k) - y_{\sigma_\theta^{T(u,X),Y}(k)}) \end{array} \right. , \quad (3)$$

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 $u^{(0)} \sim \nu$, we may now define formally the following fixed-step SGD scheme for F:

$$\begin{aligned} u^{(t+1)} &= u^{(t)} - \alpha\varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}), \\ \left(u^{(0)}, (X^{(t)})_{t\in\mathbb{N}}\ (Y^{(t)})_{t\in\mathbb{N}}\ (\theta^{(t)})_{t\in\mathbb{N}}\right) &\sim \nu \otimes \mathbb{x}^{\otimes\mathbb{N}} \otimes \mathbb{y}^{\otimes\mathbb{N}} \otimes \mathbb{\sigma}^{\otimes\mathbb{N}}. \end{aligned} \tag{4}$$

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 $\max\limits_{k\in[\![1,n]\!]} \|x_k\|_2$ 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).*

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

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

In order to deduce regularity results on f and F from Proposition 1, we will make the following assumption, which under Assumption 1 only requires additional regularity with respect to the data argument.

Assumption 2. *For any compacts $\mathcal{K}_1 \subset \mathbb{R}^{d_u}$ and $\mathcal{K}_2 \subset \mathbb{R}^{d_x}$, there exists $L_{\mathcal{K}_1,\mathcal{K}_2} > 0$ such that $\forall(u_1, u_2, x_1, x_2) \in \mathcal{K}_1^2 \times \mathcal{K}_2^2$, $\|T(u_1, x_1) - T(u_2, x_2)\| \le L_{\mathcal{K}_1,\mathcal{K}_2}(\|u_1 - u_2\| + \|x_1 - x_2\|)$.*

Proposition 2 (Regularity of $u \longmapsto f(u, X, Y, \theta)$)**.** *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 $K_f(\varepsilon, u_0, X, Y) := 2Ln(\varepsilon L + \|T(u_0, X)\|_{\infty,2} + \|Y\|_{\infty,2})$, with $L := L_{\overline{B}(u_0,\varepsilon),\overline{B}(0_{\mathbb{R}^{d_x}},\|X\|_{\infty,2})}$. Then $f(\cdot, X, Y, \theta)$ is $K_f(\varepsilon, u_0, X, Y)$-Lipschitz in $B(u_0, \varepsilon)$:*

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

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 $r := \varepsilon L_{\overline{B}(u_0,\varepsilon),\overline{B}(0_{\mathbb{R}^{d_x}},\|X\|_{\infty,2})}$.

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):

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

□

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.

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 $R_x := \sup\limits_{x\in\mathcal{X}} \|x\|_2$ and $R_y := \sup\limits_{y\in\mathcal{Y}} \|y\|_2$.*

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

Let $K_F(\varepsilon, u_0) := 2Ln(\varepsilon L + C_1(u_0) + C_2)$. We have $\forall u, u' \in B(u_0, \varepsilon)$, $|F(u) - F(u')| \le K_F(\varepsilon, u_0)\|u - u'\|_2$.

Proof. Let $\varepsilon > 0$, $u_0 \in \mathbb{R}^{d_u}$ and $^u, u' \in B(u_0, \varepsilon)$. First, notice that for any $X \in \mathcal{X}^n$, $\|X\|_{\infty,2} \leq R_x$, thus $L_{\overline{B}(u_0,\varepsilon),\overline{B}(0_{\mathbb{R}^{d_x}},\|X\|_{\infty,2})} \leq L_{\overline{B}(u_0,\varepsilon),\overline{B}(0,R_x)} =: L$. We have

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

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 $C_1(u_0) := \displaystyle\int_{\mathcal{X}^n} \|T(u_0, X)\|_{\infty,2} \mathrm{d}\mathbb{x}^{\otimes n}(X)$, which verifies $C_1(u_0) \leq M(u_0) \mathbb{x}(\mathcal{X})^n$. Now 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 $C_2 := \displaystyle\int_{\mathcal{Y}^n} \|Y\|_{\infty,2} \mathrm{d}\mathbb{y}^{\otimes n}(Y) < +\infty$.

Finally, $|F(u) - F(u')| \leq 2Ln(\varepsilon L + C_1(u_0) + C_2) \|u - u'\|_2$. $\square$

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:

$$\partial_C F(u) := \mathrm{conv} \left\{ v \in \mathbb{R}^{d_u} \ : \ \exists (u_i) \in (\mathcal{D}_F)^{\mathbb{N}} : u_i \xrightarrow[i \longrightarrow +\infty]{} u \text{ and } \nabla F(u_i) \xrightarrow[i \longrightarrow +\infty]{} v \right\}, \tag{5}$$

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.

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 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 $u(\cdot) \in \mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u})$) and start within $\mathcal{K} \subset \mathbb{R}^{d_u}$, a fixed compact set. We can now define the solution set formally as

$$S_{-\partial_C F}(\mathcal{K}) := \left\{ u \in \mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u}) \mid \underline{\forall} s \in \mathbb{R}_+, \ \dot{u}(s) \in -\partial_C F(u(s)); \ u(0) \in \mathcal{K} \right\}, \tag{6}$$

where we write $\underline{\forall}$ for "almost every". In order to compare the discrete SGD trajectories to this set continuous-time trajectories, we interpolate the discrete points in an affine manner: Equation (7) defines the *piecewise-affine interpolated SGD trajectory* associated to an SGD trajectory $(u_\alpha^{(t)})_{t \in \mathbb{N}}$ of learning rate α.

$$u_\alpha(s) = u_\alpha^{(t)} + \left(\frac{s}{\alpha} - t \right) (u_\alpha^{(t+1)} - u_\alpha^{(t)}), \quad \forall s \in [t\alpha, (t+1)\alpha[, \quad \forall t \in \mathbb{N}. \tag{7}$$

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

$$d_c(u, u') := \sum_{k \in \mathbb{N}^*} \frac{1}{2^k} \min \left(1, \max_{s \in [0,k]} \|u(s) - u'(s)\|_{\infty,2} \right). \tag{8}$$

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.

Condition 1.

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

$$\exists \varepsilon > 0,\ \forall u, u' \in B(u_0, \varepsilon),\ \forall S \in \Xi,\ |f(u, S) - f(u', S)| \leq \kappa(u_0, S)\|u - u'\|_2.$$

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

Our regularity result on f Proposition 2 allows us to verify Condition 1, by letting $\varepsilon := 1$ and $\kappa(u_0, S) := K_f(1, u_0, X, Y)$. 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 (as implied by (Tanguy et al., 2023), Lemma 2.2.2, for instance). This continuity implies that all $f(u, \cdot)$ are ζ-integrable, since $\zeta = \mathbb{x}^{\otimes n} \otimes \mathbb{y}^{\otimes n} \otimes \mathbb{\sigma}$ is a compactly supported Radon measure under Assumption 3.

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

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

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

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

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

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, yet vastly simplifies theory.

Under Assumption 4, we have for any $u \in \mathbb{R}^{d_u}$, with $\kappa(u, S) = K_f(1, u, X, Y)$ from Proposition 2 and C_2 from Proposition 3,

$$\begin{aligned}\int_{\mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y - 1}} K_f(1, u, X, Y)\mathrm{d}\mathbb{x}^{\otimes n}(X)\mathrm{d}\mathbb{y}^{\otimes n}(Y)\mathrm{d}\mathbb{\sigma}(\theta) &\leq 2Ln\left(\varepsilon L + (1 + \|u\|_2)\int_{\mathcal{X}^n} \max_{k \in [\![1, n]\!]} g(x_k)\mathrm{d}\mathbb{x}^{\otimes n}(X) + C_2\right)\\ &\leq c(1 + \|u\|_2).\end{aligned}$$

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

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

With $\lambda_{\mathbb{R}^{d_u}}$ denoting the Lebesgue measure on $\mathbb{R}^{d_u}$, let $\Gamma := \{\alpha \in]0, +\infty[\mid \forall \rho \ll \lambda_{\mathbb{R}^{d_u}},\ \rho P_\alpha \ll \lambda_{\mathbb{R}^{d_u}}\}$. Γ is the set of learning rates α for which the kernel P_α maps any absolutely continuous probability measure to another such measure. We will verify the following condition:

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 bear strong consequences on the behaviour of T for $\|u\|_2 \gg 1$, and are only practical for networks of the form $T(u,x) = \widetilde{T}(u,x)\mathbb{1}_{B(0,R)}(u)$, similarly to Assumption 4.

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

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

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

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

Let $\tau \in \mathfrak{S}_n$ and $(X, Y, \theta) \in \Xi$. Using Assumption 1, separate $I_\tau(S) = \displaystyle\sum_{j \in J} \int_{\mathcal{U}_j(X)} \mathbb{1}_B \left(u - \psi_{\tau,S}(u) \right) \mathrm{d}\rho(u)$, 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 $j = (j_1, \cdots, j_n) \in J_1(x_1) \times \cdots \times J_n(x_n)$, $\mathcal{U}_j(X) := \displaystyle\bigcap_{k=1}^n \mathcal{U}_{j_k}(x_k)$, 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 S, 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, S, τ, such that ψ is K-Lipschitz on $\mathcal{U}_j(X)$.

First, let $\chi := \left\{ \begin{array}{ccc} \mathcal{U}_j(X) & \longrightarrow & \mathbb{R}^{d_u} \\ u & \longmapsto & \left(\dfrac{\partial T}{\partial u}(u, x_k) \right)^T \theta\theta^T T(u, x_k) \end{array} \right.$. 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 $\left|\!\left|\!\left| \dfrac{\partial \chi}{\partial u}(u) \right|\!\right|\!\right|_2$. Notice that $\chi(u) = \dfrac{1}{2} \dfrac{\partial}{\partial u} (\theta \cdot T(u, x_k))^2$.

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

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

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

Second, let $\omega : u \in \mathcal{U}_j(X) \longmapsto \left(\frac{\partial T}{\partial u}(u, x_k)\right)^T \theta\theta^T y_{\tau(k)}$, also of class $\mathcal{C}^1$. We re-write $\left[\frac{\partial \omega}{\partial u}(u)\right]_{i_1,i_2} = y_{\tau(k)}^T \theta\theta^T \frac{\partial^2 T}{\partial u_{i_1} \partial u_{i_2}}(u, x_k)$, 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 $\int_{\mathcal{U}_j(X)} \mathbb{1}_B(\phi(u))\mathrm{d}\rho(u) = \int_{\mathcal{U}_j(X)} \mathbb{1}_B(u')\mathrm{d}\phi\#\rho(u') = \phi\#\rho(B)$, we have now shown that ϕ is a $\mathcal{C}^1$-diffeomorphism, thus since $\rho \ll \lambda$, $\phi\#\rho \ll \lambda$. ($\alpha \ll \beta$ denoting that α is absolutely continuous with respect to β). Since $\lambda(B) = 0$, it follows that the integral is 0, then by sum over j, $I_\tau(S) = 0$ and finally $\rho P_\alpha(B) = 0$ by integration over S and sum over τ. $\square$

Now that we have verified Condition 1, Condition 2 and Condition 3, we can apply (Bianchi et al., 2022), Theorem 2 to F. Let $\alpha_1 < \alpha_0$ (see Proposition 4).

Theorem 1 (Convergence of the interpolated SGD trajectories)**.** *Consider a neural network T and measures $\mathbb{x}$, $\mathbb{y}$ satisfying Assumption 1, Assumption 2, Assumption 3, Assumption 4 and Assumption 5.*

Let $(u_\alpha^{(t)}), \alpha \in]0, \alpha_1], t \in \mathbb{N}$ a collection of SGD trajectories associated to (4). Consider (u_α) their associated interpolations. For any compact $\mathcal{K} \subset \mathbb{R}^{d_u}$ and any $\varepsilon > 0$, we have:

$$\lim_{\substack{\alpha \longrightarrow 0 \\ \alpha \in \;]0,\alpha_1]}} \nu \otimes \mathbb{x}^{\otimes \mathbb{N}} \otimes \mathbb{y}^{\otimes \mathbb{N}} \otimes \mathbb{\sigma}^{\otimes \mathbb{N}} \left(d_c(u_\alpha, S_{-\partial_C F}(\mathcal{K})) > \varepsilon\right) = 0. \tag{9}$$

The distance d_c is defined in (8). 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 $\lambda_{\mathbb{R}^{d_u}}$-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 $R_u > 0$, 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 $\eta \ll \lambda_{\mathbb{R}^u}$, which is often taken as standard Gaussian in practice. These additions yield the following SGD scheme:

$$\begin{aligned} u^{(t+1)} = \pi_r\left(u^{(t)} - \alpha\varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}) + \alpha a\varepsilon^{(t+1)}\right), \\ \left(u^{(0)}, (X^{(t)})_{t\in\mathbb{N}}\ (Y^{(t)})_{t\in\mathbb{N}},\ (\theta^{(t)})_{t\in\mathbb{N}},\ (\varepsilon^{(t)})_{t\in\mathbb{N}}\right) \sim \nu \otimes \mathbb{x}^{\otimes\mathbb{N}} \otimes \mathbb{y}^{\otimes\mathbb{N}} \otimes \mathbb{\sigma}^{\otimes\mathbb{N}} \otimes \eta^{\otimes\mathbb{N}}, \end{aligned} \tag{10}$$

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 (10). 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 Section A.1. We formulate this condition from (Bianchi et al., 2022) before presenting sufficient conditions on T under which path differentiability shall hold.

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 $s > 0$, $\forall v \in \partial_C F(u(s))$, $v \cdot \dot{u}(s) = (F \circ u)'(s)$.*

Remark There are alternate equivalent formulations for Condition 4. Indeed, as presented in further detail in Section A.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 $\mathbb{x}$ and the data measure $\mathbb{y}$ are discrete measures, which is the case for $\mathbb{y}$ in the case of generative neural networks, but is less realistic for $\mathbb{x}$ 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 $\mathbb{x} = \sum\limits_{k=1}^{n_x} a_k\delta_{x_k}$ and $\mathbb{y} = \sum\limits_{k=1}^{n_y} b_k\delta_{y_k}$, 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 $T(\cdot, x)$ is 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 A.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.

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 $\mathcal{E} : \left\{\begin{array}{ccc} \mathbb{R}^{n\times d_y} \times \mathbb{R}^{n\times d_y} & \longrightarrow & \mathbb{R}_+ \\ Y, Y' & \longmapsto & \mathrm{SW}_2^2(\gamma_Y, \gamma_{Y'}) \end{array}\right.$. 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 (10) converge towards to a set of generalised critical points $\mathcal{Z}_r$[2] defined as

$$\mathcal{Z}_r := \left\{ u \in \mathbb{R}^{d_u} \mid 0 \in -\partial_C F(u) - \mathcal{N}_r(u) \right\}, \quad \mathcal{N}_r(u) = \left\{ \begin{array}{c} \{0\} \text{ if } \|u\|_2 < r \\ \{\lambda u \mid \lambda \geq 0\} \text{ if } \|u\|_2 = r \\ \varnothing \text{ if } \|u\|_2 > r \end{array} \right., \tag{11}$$

where $\mathcal{N}_r(u)$ refers to the *normal cone* of the ball $B(0, r)$ at x. The term $\mathcal{N}_r(u)$ in (11) 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, (11) simplifies to $\mathcal{Z}_r \cap \mathcal{D} = \{u \in \mathcal{D} \mid \nabla F(u) = 0\}$, 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.

Theorem 2 (Bianchi et al. (2022), Theorem 6 applied to (10))**.** *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 (10) for $r > 0$ and $\alpha \in]0, \alpha_1]$. One has*

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

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 $\mathcal{Z}_r$ in probability. Omitting the points of non-differentiability and the pathological case $\|u\|_2 = r$, the general idea is 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 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 A.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). As a result, we leave the generalisation to a non-discrete input measure $\mathbb{x}$ for future work.

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 (6) 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.

---

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

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`.

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.

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

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`.

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.

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.

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.

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`.

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`.

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`.

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

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`.

A 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 (5) (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.

A.1 Conservative Fields

The Clarke differential ∂_C of a locally Lipschitz function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ (defined in Equation (5)) 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 $\gamma \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p)$ 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:

$$g(u) = g(0) + \int_0^1 \max_{v \in D(\gamma(s))} \langle \dot{\gamma}(s), v \rangle \mathrm{d}s, \quad \forall \gamma \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p) \text{ such that } \gamma(0) = 1 \text{ and } \gamma(1) = u. \tag{12}$$

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:*

$$\forall u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^d),\ \underline{\forall} s > 0,\ \forall v \in \partial_C g(u(s)),\ v \cdot \dot{u}(s) = (g \circ u)'(s). \tag{13}$$

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).

A.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 seen as a generalised Jacobian matrix (see (Bolte & Pauwels, 2021), Section 3.3 for a complete presentation). $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{q\times p}$ is a conservative mapping for such a g if

$$\forall u \in \mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^p),\ \underline{\forall} s > 0,\ (g \circ u)'(s) = M\dot{u}(t),\ \forall M \in J(u(s)). \tag{14}$$

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

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

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.

A.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).