Properties of Discrete Sliced Wasserstein Losses

Eloi Tanguy, Rémi Flamary, Julie Delon

PROPERTIES OF DISCRETE SLICED WASSERSTEIN LOSSES

ELOI TANGUY, RÉMI FLAMARY, AND JULIE DELON

Abstract. The Sliced Wasserstein (SW) distance has become a popular alternative to the Wasserstein distance for comparing probability measures. Widespread applications include image processing, domain adaptation and generative modelling, where it is common to optimise some parameters in order to minimise SW, which serves as a loss function between discrete probability measures (since measures admitting densities are numerically unattainable). All these optimisation problems bear the same sub-problem, which is minimising the Sliced Wasserstein energy. In this paper we study the properties of $\mathcal{E} : Y \longmapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, i.e. the SW distance between two uniform discrete measures with the same amount of points as a function of the support $Y \in \mathbb{R}^{n\times d}$ of one of the measures. We investigate the regularity and optimisation properties of this energy, as well as its Monte-Carlo approximation $\mathcal{E}_p$ (estimating the expectation in SW using only p samples) and show convergence results on the critical points of $\mathcal{E}_p$ to those of $\mathcal{E}$, as well as an almost-sure uniform convergence and a uniform Central Limit result on the process $\mathcal{E}_p$. Finally, we show that in a certain sense, Stochastic Gradient Descent methods minimising $\mathcal{E}$ and $\mathcal{E}_p$ converge towards (Clarke) critical points of these energies.

Contents

Date: 18/03/2024.
2020 *Mathematics Subject Classification.* 49Q22 (Optimal Transport).

1. Introduction

Optimal Transport (OT) has grown in popularity as a way of lifting a notion of cost between points in a space onto a way of comparing measures on said space. In particular, endowing $\mathbb{R}^d$ with a p-norm yields the Wasserstein distance, which metrises the convergence in law on the space of Radon measures with a finite moment of order p.

The most studied object that arises from this theory is perhaps the 2-Wasserstein distance, which is defined as follows (see [40, 42, 48] for a complete practical and theoretical presentation):

$$\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{W}_2^2(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_1 - x_2\|^2 \mathrm{d}\pi(x_1, x_2), \tag{1.1}$$

where $\Pi(\mu,\nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ of first marginal μ and second marginal ν. We denote $\mathcal{P}_2(\mathbb{R}^d)$ as the set of probability measures on $\mathbb{R}^d$ admitting a second-order moment.

The 1 and 2-Wasserstein distances are commonly used for generation tasks, formulated as probability density fitting problems. One defines a statistical model μ_θ, a probability measure which is designed to approach a target data distribution μ. A typical way of solving this problem is to minimise in θ the distance between μ_θ and μ: one may choose any probability discrepancies (Kullback-Leibler, Ciszar divergences, f-divergences or Maximum Mean Discrepancy), or alternatively the Wasserstein Distance. In the case of Generative Adversarial Networks, the so-called "Wasserstein GAN" [2, 24] draws its formulation from the dual expression of the 1-Wasserstein distance.

Unfortunately, computing the Wasserstein distance is prohibitively costly in practice. The discrete formulation of the Wasserstein distance (the Kantorovich linear problem) is typically solved approximately using standard linear programming tools. These methods suffer from a super-cubic worst-case complexity with respect to the number of samples from the two measures. Furthermore, given n samples from each measure μ and ν, the convergence of the estimated distance $\mathrm{W}_2(\hat{\mu}_n, \hat{\nu}_n)$ is only in $\mathcal{O}(n^{-1/d})$ towards the true distance, thus OT suffers from the curse of dimensionality, as is known since Dudley, 1969 [20].

Several efforts have been made in recent years to make Optimal Transport more accessible computationally. In particular, many surrogates for W_2 have been proposed, perhaps the most notable of which is the Sinkhorn Divergence (see [40, 15, 23]). The Sinkhorn Divergence adds entropic regularisation to OT, yielding a strongly convex algorithm which can be solved efficiently.

Another alternative is the Sliced Wasserstein (SW) Distance, which leverages the simplicity of computing the Wasserstein distance between one-dimensional measures. Indeed, given

$$\gamma_X := \frac{1}{n}\sum_{k=1}^{n} \delta_{x_k},\ \gamma_Y := \frac{1}{n}\sum_{k=1}^{n} \delta_{y_k} \text{ with } x_1, \cdots, x_n, y_1, \cdots, y_n \in \mathbb{R},$$

the 2-Wasserstein distance between these two measures can be computed by sorting their supports:

$$\mathrm{W}_2^2(\gamma_X, \gamma_Y) = \frac{1}{n}\sum_{k=1}^{n} (x_{\sigma(k)} - y_{\tau(k)})^2, \tag{1.2}$$

where σ is a permutation sorting $(x_1, \cdots, x_n)$, and τ is a permutation sorting $(y_1, \cdots, y_n)$.

The idea of the Sliced Wasserstein Distance [41] is to compute the 1D Wasserstein distances between projections of input measures. We write $P_\theta : \mathbb{R}^d \longrightarrow \mathbb{R}$ the map $x \longmapsto \theta^T x$, and ϖ the uniform measure over the euclidean unit sphere of $\mathbb{R}^d$, $\mathbb{S}^{d-1}$. Denoting $\#$ the push-forward operation [1], the Sliced Wasserstein distance between two measures μ and ν is defined as

$$\mathrm{SW}_2^2(\mu, \nu) := \int_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \mu, P_\theta \# \nu) \mathrm{d}\varpi(\theta). \tag{1.3}$$

Similarly, for $q \geq 1$, the q-Sliced Wasserstein distance SW_q^q (to the power q) is obtained by replacing W_2^2 in the previous equation by the q Wasserstein distance (to the power q) W_q^q.

SW has enjoyed a substantial amount of theoretical study, albeit not as extensively as for the original Wasserstein distance. For measures supported on a fixed compact of $\mathbb{R}^d$, Bonnotte ([13], Chapter 5) has shown that the Wasserstein and Sliced Wasserstein distances are equivalent. The same work also developed a theory of gradient flows for SW, which justifies some generative methods. Further discussion on this equivalence has been performed by Bayraktar and Guo [5]. Nadjahi et al. [38] showed that SW metrises the convergence in law (without restrictions of the measure supports), and further concluded guarantees for SW-based generative models.

Continuous measures being out of the reach of practical computation, it is necessary to perform sample estimation and replace them with discrete empirical estimates. Thankfully, as shown in [37], the *sample complexity* (i.e. the rate of convergence of the estimates w.r.t. the number of samples) for sliced distances such as SW is in $1/\sqrt{n}$, which in particular avoids the curse of dimensionality from which the Wasserstein Distance suffers. This fuels interest for the study of $Y \mapsto \mathrm{SW}(\gamma_Y, \gamma)$, which is to say the variation of SW w.r.t. the discrete support

[1] The push-forward of a measure μ on $\mathbb{R}^d$ by an application $T : \mathbb{R}^d \to \mathbb{R}^k$ is defined as a measure $T\#\mu$ on $\mathbb{R}^k$ such that for all Borel sets $B \in \mathcal{B}(\mathbb{R}^k), T\#\mu(B) = \mu(T^{-1}(B))$.

of one of the measures. It is currently unknown whether this functional presents strict local optima, for instance.

Originally, SW was introduced as a more computable alternative to the Wasserstein distance, notably for texture mixing using a barycentric formulation [41, 11]. Other uses of SW have been suggested, notably in statistics as a probability discrepancy. For instance, Nadjahi et al. [36] proposed an approximate bayesian computation method, where the estimation of the posterior parameters is done by selecting those under which the SW distance between observed and synthetic data is below a fixed threshold. Other widespread uses of SW in image processing include colour transfer [1] and colour harmonisation [12].

Nowadays, SW is commonly used as a training or validation loss in generative Machine Learning. Karras et al. [29] propose to use SW to compare GAN results, by comparing images via multi-scale patched descriptors. Some generative models (including GANs and auto-encoders), leverage the computational advantages of SW in order to learn a target distribution. This is done under the implicit generative modelling framework, where a network T_u of parameters u is learned such as to minimise $\mathrm{SW}(T_u\#\mu_0, \mu)$, where μ_0 is a low-dimensional input distribution (often chosen as Gaussian or uniform noise), and where μ is the target distribution. Deshpande et al. [19] and Wu et al. [50] train GANs and auto-encoders within this framework; Liutkus et al. [33] perform generation by minimising a regularised SW problem, which they solve by gradient flow using an SDE formulation. SW can be used to synthesise images by minimising the SW distance between features of the optimised image and a target image, as done by [26] for textures with neural features, and by [44] with wavelet features (amongst other methods).

In practice, the integration over the unit sphere in SW is intractable, and one must resort to a Monte-Carlo approximation, taking the average between p projections instead of the expectation, usually during iterations of a Stochastic Gradient Descent [30]. This implies that for a finite number of iterations, a fixed number of projections p, potentially very small compared to what is needed to explore the hypersphere, is explored in practice. The question of this finite number of final projection directions is made even more important by the fact that practitioners usually optimise the expectation of the SW distance on large mini-batches [19] that also limits the total number of effective projections p. The estimation error of this approximation has not been extensively studied, and it is common in practice to assume that this empirical version presents the same properties as the true SW distance.

An important question is the conditions under which these approximations for SW are valid. In practice, sliced-Wasserstein Generative Models compute SW in the data space or in the data encoding space ([30, 19]), which yields high values for the dimension d, in particular for images. Note that the necessity behind having a large number of projections p was already hinted at in [30], §3.3. Another untreated question is the complexity of optimising this approximation of SW, and how this optimisation landscape compares to the true SW landscape.

Bonneel et al. [11] studied the uses of SW for barycentre computation, and in particular proved that the empirical SW distance is $\mathcal{C}^1$ on a certain open set, with respect to the measure positions. They remarked that in practice, numerical resolutions for discretised SW distances converged towards (eventual) local optima, however the convergence and local optima have not been studied theoretically.

In this paper, we propose to study $\mathcal{E} : Y \mapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, where γ_Y and γ_Z are two uniform discrete measures supported by n points, denoted by Y and Z. Our main objective is to provide optimisation properties for the landscapes of $\mathcal{E}$ and its Monte-Carlo counterpart $\mathcal{E}_p$, obtained by replacing the expectation by an average over p projections. In Section 2, we prove several regularity properties for both energies, such as semi-concavity, and we show that the convergence of the Monte-Carlo estimation is uniform (on every compact) w.r.t. the measure locations. Section 3 focuses on the respective landscapes of $\mathcal{E}$ and $\mathcal{E}_p$, and shows that the critical points of $\mathcal{E}$ satisfy a fixed-point equation, and how the critical points of $\mathcal{E}_p$ relate to this fixed-point equation when the number of projections p increases (with convergence rates). Mérigot et al. follow a similar methodology in [35], where they study optimisation properties for $Y \longmapsto \mathrm{W}_2(\gamma_Y, \mu)$, with μ a continuous measure. The main difficulty they face arises from the non-convexity of the map, and this difficulty is also central in our work. The last two sections of our paper tackle numerical considerations. To begin with, since $\mathcal{E}$ and $\mathcal{E}_p$ are usually minimised in the literature using Stochastic Gradient Descent (SGD), we provide in Section 4 the first complete convergence study of SGD for $\mathcal{E}$ and $\mathcal{E}_p$, relying on the recent works [6] and [18]. Finally, Section 5 challenges our theoretical results with extensive numerical experiments, quantifying the impact of the dimension and several other parameters on the convergence.

Notations.

- d is the dimension, n is the number of points
- p is the number of projections $(\theta_1, \cdots, \theta_p)$
- $\|\cdot\|_2$: Euclidean norm of $\mathbb{R}^n$
- Matrices $X \in \mathbb{R}^{n\times d}$ are written $X = (x_1, \cdots, x_n)^T$ with the $x_i \in \mathbb{R}^d$
- $\|Y\|_{\infty,2}$ for $Y \in \mathbb{R}^{n\times d}$ denotes $\max_{i\in[\![1,n]\!]} \|Y_{i,\cdot}\|_2 = \max_{i\in[\![1,n]\!]} \|y_i\|_2$
- $M \cdot N$: inner product $\mathrm{Trace}(M^T N)$ for matrices
- W_2: 2-Wasserstein Distance (1.1)
- ϖ: Uniform measure on the unit sphere $\mathbb{S}^{d-1}$ of $\mathbb{R}^d$
- P_θ: for $\theta \in \mathbb{S}^{d-1}$, $P_\theta = x \longmapsto \theta^T x$
- SW_2: Sliced 2-Wasserstein distance (1.3)
- SW_q: Sliced q-Wasserstein distance
- γ_X: for $X \in \mathbb{R}^{n\times d}$: discrete measure $\frac{1}{n}\sum_i \delta_{X_{i,\cdot}} = \frac{1}{n}\sum_i \delta_{x_i}$
- $\mathcal{E}(Y)$: $\mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$ (2.1)
- $\mathcal{E}_p(Y)$: Monte-Carlo approximation of $\mathcal{E}(Y)$ with p projections (2.2)
- Σ_n: n-simplex: $a \in \mathbb{R}_+^n$ such that $\sum_i a_i = 1$
- $\|M\|_F$: Frobenius norm: $\sqrt{\sum_{i,j} M_{i,j}^2}$
- $\mathbf{m}$ denotes p permutations $(\sigma_1, \cdots, \sigma_p)$ of $[\![1, n]\!]$, see Section 2.3
- $\mathcal{C}_{\mathbf{m}}$: cell of configuration $\mathbf{m}$, see Section 2.3

2. Sliced and Empirical Sliced Wasserstein Energies and their Regularities

2.1. **The discrete SW energies $\mathcal{E}$ and $\mathcal{E}_p$.** The Sliced Wasserstein distance has been widely studied as an alternative to the Wasserstein distance, in particular it is arguably simpler to compute in order to minimise measure discrepancies. In practice, one may not work with continuous measures, which are beyond the capabilities of numerical approximations, thus one must sometimes contend with discrete

measures. To that end, we study in this paper the SW distance between discrete measures, and in particular the associated energy landscape with respect to the support of one of the measures:

$$\mathcal{E} := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ Y & \longmapsto & \displaystyle\int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z) \mathrm{d}\varpi(\theta) \end{array} \right. , \tag{2.1}$$

where n denotes the number of points in the data matrices Y, Z, which we write as data entries stacked vertically: $Y = (y_1, \cdots, y_n)^T$, with points in $\mathbb{R}^d$. The associated (uniform) discrete measure supported on $\{y_1, \cdots, y_n\}$ will be denoted $\gamma_Y := \frac{1}{n}\sum_k \delta_{y_k}$.

For instance, this framework encompasses SW-based implicit generative models ([19], [50]), which optimise parameters ρ by minimising $\mathrm{SW}(T_\rho \# \mu_0, \mu)$, where μ_0 is comprised of samples of a simple distribution, and μ corresponds to data samples which we would like to generate. In this setting, one would need to minimise *through* $\mathcal{E}$. The use of discrete measures is also backed theoretically by the study of the *sample complexity* of SW [37], which is to say the rate of decrease of the approximation error between $\mathrm{SW}(\mu, \nu)$ and its discretised counterpart $\mathrm{SW}(\hat{\mu}_n, \hat{\nu}_n)$.

In practical and realistic settings, the only numerically accessible workaround to optimise through $\mathcal{E}$ is a form of discretisation of the set of directions. The first and most common method, due to its efficiency and simplicity, is to minimize $\mathcal{E}$ through stochastic gradient descent (SGD): at each time set t, p random directions $(\theta_1^{(t)}, \cdots, \theta_p^{(t)})$ are drawn, and a gradient descent step is performed by approximating $\mathcal{E}$ by a discrete sum on these p random directions. This method is optimisation-centric, since it does not concern itself with computing the final SW distance and focuses on optimising the parameters. A second possible discretisation method consists in fixing the p directions $(\theta_1, \cdots, \theta_p)$ once for all and replacing $\mathcal{E}$ in the minimization by its Monte-Carlo estimator [2]

$$\mathcal{E}_p := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ Y & \longmapsto & \displaystyle\frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(P_{\theta_i} \# \gamma_Y . P_{\theta_i} \# \gamma_Z) \end{array} \right. . \tag{2.2}$$

It is important to note that both methods are intuitively tied, since in both cases there is a finite amount of sampled directions. If the SGD method lasts T iterations with p projections every time, it amounts to a specific way of optimising $\mathcal{E}_{pT}$. For this reason, studying $\mathcal{E}_p$ theoretically is not only interesting in itself as an approximation of $\mathcal{E}$, but also yields a better insight on the SGD strategy.

The study of $\mathcal{E}$ is also tied with the study of the SW barycentres, which solve the optimisation problem

$$\mathrm{Bar}(\lambda_j, \gamma_{Z^{(j)}})_{j\in[\![1,J]\!]} = \operatorname*{argmin}_{Y\in\mathbb{R}^{n\times d}} \sum_{j=1}^{J} \lambda_j \mathcal{E}(Y, Z^{(j)}) =: \mathcal{E}_{\mathrm{bar}}(Y), \tag{2.3}$$

where the notation $\mathcal{E}(Y, Z^{(j)})$ reflects the dependency on Z in the definition of $\mathcal{E}$ (2.1). The regularity and convergence results will immediately be applicable to the barycentre energy (2.3). While the optimisation results on $\mathcal{E}$ and $\mathcal{E}_p$ will not generalise naturally due to the sum, the SGD convergence results shall still hold.

[2] In this notation the projection axes $\theta_1, \cdots, \theta_p \in \mathbb{S}^{d-1}$ are written implicitly, the complete notation being $\mathcal{E}_p(Y; (\theta_i)_{i\in[\![1,p]\!]})$ when required.

As a Monte-Carlo estimator, the law of large numbers yields the point-wise convergence of $\mathcal{E}_p$ to $\mathcal{E}$ if the $(\theta_i)_{i\in\mathbb{N}}$ are i.i.d. of law ϖ:

$$\mathcal{E}_p(Y;(\theta_i)_{i\in[\![1,p]\!]}) \xrightarrow[p\to+\infty]{\text{a.s.}} \mathcal{E}(Y). \tag{2.4}$$

For this reason, it is often assumed that numerically, $\mathcal{E}_p$ and $\mathcal{E}$ will behave similarly, which is perhaps why research has been scarce on the landscape of $\mathcal{E}_p$, the focus remaining on the theoretical properties of the true or mini-batch Sliced Wasserstein Distance [38, 36]. But as discussed in the introduction, practitioners often optimize the SW distance using SGD with a finite number of projection directions [30, 19], and the landscape of $\mathcal{E}_p$ is of paramount importance. This section and the next one are dedicated to studying the relations and differences between $\mathcal{E}_p$ and $\mathcal{E}$.

Remark 2.1. *Some of our results can in fact be extended to q-SW instead of 2-SW, especially regularity results Lemma 2.1, Proposition 2.1 and Theorem 2.1, as well as the statistical estimation results Theorem 2.3 and Theorem 2.4. However, as soon as we need the cell structure of $\mathcal{E}_p$ (Section 2.3), we leverage the simplicity of the quadratic case $q = 2$.*

2.2. **Regularity properties of $\mathcal{E}_p$ and $\mathcal{E}$.** In order to study the regularity of our energies, we first focus on the regularity of w_θ, the 2-Wasserstein distance between two discrete measures projected on the line $\mathbb{R}\theta$:

$$w_\theta := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R} \\ Y & \longmapsto & \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z) \end{array} \right. . \tag{2.5}$$

With this notation, observe that $\mathcal{E}$ and $\mathcal{E}_p$ can be written

$$\mathcal{E}(Y) = \mathbb{E}_{\theta\sim\varpi}\left[w_\theta(Y)\right] \quad \text{and} \quad \mathcal{E}_p(Y) = \mathbb{E}_{\theta\sim\varpi_p}\left[w_\theta(Y)\right], \tag{2.6}$$

where $\varpi_p := \dfrac{1}{p}\displaystyle\sum_{i=1}^{p} \delta_{\theta_i}$ for p fixed directions $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$.

We now provide an important regularity result about the uniformly locally Lipschitz property of the functions $(w_\theta)_\theta$, which will yield easily that our energies $\mathcal{E}$ and $\mathcal{E}_p$ are also locally Lipschitz, a central property in the convergence study of particular SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$ (see Section 4.2). To show this result on (w_θ), we need the following Lemma 2.1, whose proof is provided in Section A.3. This result shows that the Wasserstein cost is regular in some sense with respect to the measure weights and the cost matrix, which will be helpful when studying the regularity of the functions w_θ.

Lemma 2.1 (Stability of the Wasserstein cost)**.** *Let $\alpha, \overline{\alpha}, \beta, \overline{\beta} \in \Sigma_n$, and $C, \overline{C} \in \mathbb{R}_+^{n\times n}$. Denote by $\mathrm{W}(\alpha,\beta;C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi\cdot C$ the cost of the discrete Kantorovich problem of cost matrix C between the weights α, β. We have the following Lipschitz-like inequalities, assuming $\alpha, \overline{\alpha}, \beta, \overline{\beta} > 0$ entry-wise:*

$$\left|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};\overline{C})\right| \leq \|C-\overline{C}\|_\infty + \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1), \tag{2.7}$$

$$\left|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};\overline{C})\right| \leq \|C-\overline{C}\|_F + \|C\|_F\left(\|\alpha-\overline{\alpha}\|_2 + \|\beta-\overline{\beta}\|_2\right). \tag{2.8}$$

Remark 2.2. *Using (2.8) twice with $(C,\overline{C})$ and $(\overline{C},C)$ yields a symmetric second term with a factor $\min(\|C\|_\infty, \|\overline{C}\|_\infty)$ instead of $\|C\|_\infty$, and likewise for $\|\cdot\|_F$ with (2.8).*

Remark 2.3. *The result of Lemma 2.1 assumes positive weights, but in the case of the q-Wasserstein cost $C_{i,j} = \|x_i - y_j\|_2^q$ with $q \geq 1$, we can remove this assumption by a continuity argument, since the q-Wasserstein distance metrises the weak convergence of measures (see* [42]*, Theorem 5.10 or 5.11, applied to the simple case of discrete measures for which convergence of moments is immediate).*

The following regularity property on (w_θ) uses the norm $\|X\|_{\infty,2} = \max_{k \in [\![1,n]\!]} \|x_k\|_2$ on $\mathbb{R}^{n\times d}$. We also denote $D := n \times d$ for convenience.

Proposition 2.1. *The $(w_\theta)_{\theta\in\mathbb{S}^{d-1}}$ are uniformly locally Lipschitz.. More precisely, in a neighbourhood $X \in \mathbb{R}^D$ or radius $r > 0$, writing $\kappa_r(X) := 2n(r + \|X\|_{\infty,2} + \|Z\|_{\infty,2})$, each w_θ is $\kappa_r(X)$ Lipschitz, which is to say*

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

Proof. Let $X \in \mathbb{R}^D,\ Y, Y' \in B_{\|\cdot\|_{\infty,2}}(X, r)$, and $\theta \in \mathbb{S}^{d-1}$. By Lemma 2.1 Equation (2.8), we have $|w_\theta(Y) - w_\theta(Y')| \leq \|C - C'\|_F$, where for $(k, l) \in [\![1, n]\!]^2,\ C_{k,l} := (\theta^T y_k - \theta^T z_l)^2$, likewise for C'. Then:

$$\begin{aligned}
[C - C']_{k,l} &= \left(\theta^T(y_k - y'_k)\right)\left(\theta^T(y_k + y'_k - 2z_l)\right) \\
&\leq \|y_k - y'_k\|_2 \|y_k + y'_k - 2z_l\|_2 \\
&= \|y_k - y'_k\|_2 \|y_k - x_k + y'_k - x_k + 2z_l + 2x_k\|_2 \\
&\leq \|y_k - y'_k\|_2 \left(2r + 2\|Z\|_{\infty,2} + 2\|X\|_{\infty,2}\right).
\end{aligned}$$

Finally, $\|C - C'\|_F = \sqrt{\sum_{k,l\in[\![1,n]\!]} [C - C']_{k,l}^2} \leq 2n(r + \|X\|_{\infty,2} + \|Z\|_{\infty,2})\|Y - Y'\|_{\infty,2}$. □

As a consequence, we deduce immediately that $\mathcal{E}_p$ and $\mathcal{E}$ are locally Lipschitz.

Theorem 2.1. *$\mathcal{E}$ and $\mathcal{E}_p$ are locally Lipschitz.*

Proof. Let $X \in \mathbb{R}^D,\ r > 0$ and $\mu \in \{\mathfrak{w}, \mathfrak{w}_p\}$. By Proposition 2.1, for any $Y, Y' \in B_{\|\cdot\|_{2,\infty}}(X, r)$,

$$|\mathbb{E}_{\theta\sim\mu}[w_\theta(Y)] - \mathbb{E}_{\theta\sim\mu}[w_\theta(Y')]| \leq \mathbb{E}_{\theta\sim\mu}[|w_\theta(Y) - w_\theta(Y')|] \leq \kappa_r(X)\|Y - Y'\|_{\infty,2}.$$

□

As a locally Lipschitz function, $\mathcal{E}$ is differentiable almost everywhere. The expression of its gradient is quite simple and corresponds to the simple differentiation of w_θ in the integral, as was shown in [11]. We remind here their result for the sake of completeness, and because the derivative will be useful on several occasions in this paper. We define $\mathcal{U}$ the open set of matrices with distinct lines

$$\mathcal{U} = \left\{(x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d} \mid \forall i \neq j,\ [\![1, n]\!]^2,\ x_i \neq x_j\right\}. \tag{2.9}$$

Theorem 2.2 (Regularity of $\mathcal{E}$, from Bonneel et al. [11] Theorem 1)**.** *$\mathcal{E}$ is continuous on $\mathbb{R}^{n\times d}$, and of class $\mathcal{C}^1$ on $\mathcal{U}$. There exists $\kappa \geq 1$ such that $\nabla\mathcal{E}$ is κ-Lipschitz on $\mathcal{U}$. For $Y \in \mathcal{U}$, one has the expression:*

$$\frac{\partial\mathcal{E}}{\partial y_k}(Y) = \frac{2}{n}\int_{\mathbb{S}^{d-1}} \theta\theta^T(y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)})\mathrm{d}\mathfrak{w}(\theta), \tag{2.10}$$

where for $\theta \in \mathbb{S}^{d-1}, X \in \mathcal{U},\ \tau_X^\theta \in \mathfrak{S}_n$ is any permutations s.t. $\theta^T x_{\tau_X^\theta(1)} \leq \cdots \leq \theta^T x_{\tau_X^\theta(n)}$.

Proving this theorem requires to be cautious. Firstly, differentiating directly under the integral using standard calculus theorems is impossible, since the integrand is only differentiable on a set $\mathcal{U}_\theta$ which depends on the integration variable θ. Fortunately, these irregularities are smoothed out as θ rotates, yielding differentiability almost-everywhere. Secondly, the problematic term τ_Y^θ can be dealt with for $Y \in \mathcal{U}$ by remarking that for any Y' ε-close to Y, we have $\tau_Y^\theta = \tau_{Y'}^\theta$ for every θ in a certain subset of $\mathbb{S}^{d-1}$ which is of ϖ-measure exceeding $1 - C\varepsilon$. Regarding the multiplicative constant, Theorem 1 in Bonneel et. al omits the $1/n$ factor (we believe that this is a typing error).

2.3. **Cell structure of $\mathcal{E}_p$.** In order to further study the optimisation properties of $\mathcal{E}_p$ and $\mathcal{E}$, we need to exhibit more explicitly the structure of the landscape of $\mathcal{E}_p$. The semi-concavity of $\mathcal{E}_p$ and $\mathcal{E}$ will follow, as well as the fact that $\mathcal{E}$ is semi-algebraic (see Proposition 2.5). We can compute $\mathcal{E}_p$ by leveraging the formula for 1D Wasserstein distances:

$$\forall Y \in \mathbb{R}^{n\times d}, \quad \mathcal{E}_p(Y) = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}\left(\theta_i^T\left(y_k - z_{\tau_Z^{\theta_i}\circ(\tau_Y^{\theta_i})^{-1}(k)}\right)\right)^2. \tag{2.11}$$

For now we consider Z and the (θ_i) fixed, and we write $\mathbf{m}(Y) := (\mathbf{m}_i(Y))_{i\in[\![1,p]\!]}$ where $\mathbf{m}_i(Y) = \tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1}$. Writing $\mathfrak{S}_n$ the set of permutations of $\{1, \cdots, n\}$, $\mathbf{m}_i$ is the element σ of $\mathfrak{S}_n$ which solves the (Monge) quadratic optimal transport between the points $(\theta_i^T y_1, \cdots, \theta_i^T y_n)$ and $(\theta_i^T z_1, \cdots, \theta_i^T z_n)$. The matching configuration $\mathbf{m}(Y)$ depends implicitly on the fixed directions (θ_i).

Note that the permutations τ_Y^θ and τ_Z^θ are not always uniquely defined: for any $\theta \in \mathbb{S}^{d-1}$, there exists $Y \in \mathcal{U}$ such that τ_Y^θ is not uniquely defined (take Y such that $\theta \in (y_1 - y_2)^\perp$ for instance). However, for a given set of directions (θ_i), these permutations are uniquely defined almost everywhere on $\mathbb{R}^{n\times d}$.

A set of interest is $\mathcal{C}_\mathbf{m} = \{Y \in \mathcal{U} \mid \mathbf{m}(Y) \text{ is uniquely defined and equal to } \mathbf{m}\}$, the cell of points Y of configuration $\mathbf{m}$. Writing $\mathbf{m} = (\mathbf{m}_1, \cdots, \mathbf{m}_p)$, and using the optimality of each $\mathbf{m}_i$, note that each cell $\mathcal{C}_\mathbf{m}$ can be also written as

$$\begin{aligned}\mathcal{C}_\mathbf{m} = \Bigg\{ & Y \in \mathbb{R}^{n\times d} : \forall i \in [\![1,p]\!],\ \forall \sigma \in \mathfrak{S}_n \setminus \{\mathbf{m}_i\}, \\ & \sum_{k=1}^{n} z_{\mathbf{m}_i(k)}^T \theta_i \theta_i^T y_k > \sum_{k=1}^{n} z_{\sigma(k)}^T \theta_i \theta_i^T y_k \Bigg\}.\end{aligned} \tag{2.12}$$

Thus, each $\mathcal{C}_\mathbf{m}$ is an open polyhedral cone, obtained as the intersection of $p(n!-1)$ half-open planes. Moreover, the union of these cells $\bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \mathcal{C}_\mathbf{m}$ is a strict subset of $\mathcal{U}$ (as a consequence of the non uniqueness of the permutations for some Y), but is dense in $\mathbb{R}^{n\times d}$. These cells are of particular interest since by definition, $\mathcal{E}_p$ is quadratic on each $\mathcal{C}_\mathbf{m}$, and can be written

$$\forall Y \in \mathcal{C}_\mathbf{m},\ \mathcal{E}_p(Y) = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}\left(\theta_i^T\left(y_k - z_{\mathbf{m}_i(k)}\right)\right)^2 =: q_\mathbf{m}(Y). \tag{2.13}$$

Furthermore, the sorting interpretation of the 1D Wasserstein distance allows us to re-write $\mathcal{E}_p(Y)$ as a minimum of quadratics,

$$\forall Y \in \mathbb{R}^{n\times d},\ \mathcal{E}_p(Y) = \min_{\mathbf{m}\in\mathfrak{S}_n^p} q_{\mathbf{m}}(Y) = q_{\mathbf{m}(Y)}(Y). \tag{2.14}$$

Remark 2.4. *To each $Y = (y_1, \cdots, y_n)^T$ (seen as a $n\times d$ matrix), we can associate the column vector $\mathrm{vec}(y) := (y_1^T, \cdots, y_n^T)^T$, which is now a vector of $\mathbb{R}^D = \mathbb{R}^{n\times d}$ without any abuse of notation. We re-write the quadratic from equation (2.13) in standard form: $q_{\mathbf{m}}(\mathrm{vec}(y)) = \frac{1}{2}\mathrm{vec}(y)^T B \mathrm{vec}(y) - a_{\mathbf{m}}^T \mathrm{vec}(y) + b$, where:*

$$B := \frac{2}{n}\begin{pmatrix} A & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & A \end{pmatrix};\ A := \frac{1}{p}\sum_{i=1}^p \theta_i\theta_i^T;$$

$$a_{\mathbf{m}} := \frac{2}{pn}\begin{pmatrix} \sum_{i=1}^p \theta_i\theta_i^T z_{\mathbf{m}_i(1)} \\ \vdots \\ \sum_{i=1}^p \theta_i\theta_i^T z_{\mathbf{m}_i(n)} \end{pmatrix};\ b := \frac{1}{n}\sum_{k=1}^n z_k^T A z_k. \tag{2.15}$$

Note in particular that only the linear component depends on $\mathbf{m}$.

Finding the minimum of each quadratic $q_{\mathbf{m}}$ can be done in closed form, thanks to the computations of Remark 2.4. This computational accessibility will be leveraged during our discussions on minimising $Y \longmapsto \mathcal{E}_p(Y)$ (Section 3.2.4), wherein we shall present the Block Coordinate Descent method (Algorithm 1), which computes iteratively minima of quadratics in closed form.

2.4. **Consequences of the cell structure on the regularity of $\mathcal{E}_p$ and $\mathcal{E}$.** The cell decomposition presented in Section 2.3 permits to show several additional regularity results.

Proposition 2.2. *$\mathcal{E}_p$ is quadratic on each cell $\mathcal{C}_{\mathbf{m}}$, thus of class $\mathcal{C}^\infty$ on $\bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \mathcal{C}_{\mathbf{m}}$, hence $\mathcal{C}^\infty$ a.e..*

The formulation as an infimum of quadratics also allows us to prove that $\mathcal{E}_p$ is semi-concave, which is an extremely useful property for optimisation.

Proposition 2.3. *$\mathcal{E}_p$ is $\frac{1}{n}$-semi-concave, i.e. $\mathcal{E}_p - \frac{1}{n}\|\cdot\|_2^2$ is concave.*

Proof. Using the notations from Remark 2.4, $\mathcal{E}_p(\mathrm{vec}(y)) = \frac{1}{2}\mathrm{vec}(y)^T B\mathrm{vec}(y) + \min_{\mathbf{m}\in\mathfrak{S}_n^p} a_{\mathbf{m}}^T\mathrm{vec}(y) + b$. Now, $\mathrm{vec}(y) \longmapsto \min_{\mathbf{m}\in\mathfrak{S}_n^p} a_{\mathbf{m}}^T\mathrm{vec}(y) + b$ is concave, as an infimum of affine functions. Furthermore

$$\frac{1}{2}\mathrm{vec}(y)^T B\mathrm{vec}(y) - \frac{1}{n}\|\mathrm{vec}(y)\|_2^2 = \frac{1}{n}\sum_{k=1}^n y_k^T(A - I)y_k,$$

and since $A \preceq I_d$, the equation above defines a concave function of $\mathrm{vec}(y)$. □

The semi-concavity of $\mathcal{E}_p$ and point-wise convergence allows us to deduce the semi-concavity of $\mathcal{E}$:

Proposition 2.4. *$\mathcal{E}$ is $\frac{1}{n}$-semi-concave.*

Proof. By Proposition 2.3, $\forall p \in \mathbb{N}^*$, $\mathcal{E}_p$ is $\frac{1}{n}$-semi-concave. Let $p \in \mathbb{N}^*$, $Y, Y' \in \mathbb{R}^{n\times d}$ and $\lambda \in [0,1]$. We have

$$\begin{aligned}&\mathcal{E}_p((1-\lambda)Y + \lambda Y') - \frac{1}{n}\|(1-\lambda)Y + \lambda Y'\|_F^2\\ \geq\ &(1-\lambda)\mathcal{E}_p(Y) + \lambda\mathcal{E}_p(Y') - \frac{1}{n}\left((1-\lambda)\|Y\|_F^2 + \lambda\|Y'\|_F^2\right).\end{aligned}$$

Taking the limit $p \longrightarrow +\infty$ in this inequality yields the $\frac{1}{n}$-semi-concavity of $\mathcal{E}$. □

The cell formulation also allows us to show that $\mathcal{E}_p$ is semi-algebraic, which means that it can be written using a finite number of polynomial expressions. This result induces strong optimisation results akin to semi-concavity for our purposes. We recall the definition of a semi-algebraic set ([49], Definition 1). $S \subset \mathbb{R}^D$ is *semi-algebraic* if it can be written $S = \bigcup_{n=1}^{N}\bigcap_{m=1}^{M} A_{n,m}$ where $(A_{n,m})_{\substack{n\in[\![1,N]\!]\\ m\in[\![1,M]\!]}}$ is a finite family of sets such that $A_{n,m} = \{X \in \mathbb{R}^D \mid p_{n,m}(X) \geq 0\}$ or $A_{n,m} = \{X \in \mathbb{R}^D \mid p_{n,m}(X) = 0\}$, with $p_{n,m}$ being D-variate polynomials with real coefficients. A *semi-algebraic* function is a function whose graph is a semi-algebraic set.

Proposition 2.5. *$\mathcal{E}_p$ is semi-algebraic.*

Proof. We shall prove that the set $G := \{(X, \mathcal{E}_p(X)) \mid X \in \mathcal{U}_\Theta\}$ is semi-algebraic, where $\mathcal{U}_\Theta := \bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \mathcal{C}_\mathbf{m}$. Observe that

$$G = \bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \left\{(X,y) \in \mathbb{R}^{D+1}, X \in \mathcal{C}_\mathbf{m} \text{ and } y = q_\mathbf{m}(X)\right\}.$$

The function $q_\mathbf{m}$ is quadratic, thus polynomial, and the cells $\mathcal{C}_\mathbf{m}$ are intersections of a finite number of half planes, so we conclude that G is semi-algebraic.

The closure of $\mathcal{U}_\Theta$ verifies $\overline{\mathcal{U}_\Theta} = \mathbb{R}^D$, furthermore, since $\mathcal{E}_p$ is continuous on $\mathbb{R}^D$ (by Theorem 2.2), the closure of G is exactly the graph of $\mathcal{E}_p$. Now by [49], Lemma 4, since G is semi-algebraic, then $\overline{G}$ is also semi-algebraic. As a conclusion, $\mathcal{E}_p$ is a semi-algebraic function. □

2.5. **Convergence of $\mathcal{E}_p$ to $\mathcal{E}$.** We have already seen that $\mathcal{E}_p(Y)$ converges to $\mathcal{E}(Y)$ almost surely when $p \to +\infty$. In practice, since we want to optimise through $\mathcal{E}_p$ as a surrogate for $\mathcal{E}$, we would wish for the strongest possible convergence. Below, we show almost-sure *uniform* convergence over any compact, which is substantially better than point-wise convergence. Note that this stronger mode of convergence is unfortunately still too weak to transport local optima properties.

Theorem 2.3 (Uniform Convergence of $\mathcal{E}_p$)**.** *Let $\mathcal{K} \subset \mathbb{R}^{n\times d}$ compact. We have*

$$\mathbb{P}\left(\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\to+\infty]{} 0\right) = 1, \text{ where for } f \in \mathcal{C}(\mathcal{K},\mathbb{R}),\ \|f\|_{\ell^\infty(\mathcal{K})} := \sup_{x\in\mathcal{K}} |f(x)|.$$

Proof. We shall temporarily write $\mathcal{E}_p(Y) = \mathcal{E}_p(Y;\Theta)$ to illustrate the dependency on the random variable $\Theta := (\theta_i)_{i\in\mathbb{N}^*}$ on a probabilistic space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(\mathbb{S}^{d-1})^{\mathbb{N}}$. By point-wise almost-sure convergence, for any fixed $Y \in \mathbb{R}^{n\times d}$, there exists a $\mathbb{P}$-null set $\mathcal{N}_Y$ such that for every $\omega \in \Omega \setminus \mathcal{N}_Y$, the deterministic real

number $\mathcal{E}_p(Y;\Theta(\omega))$ converges to $\mathcal{E}(Y)$. Let $\mathcal{D} := \mathcal{K} \cap \mathbb{Q}^{n\times d}$, which is dense in $\mathcal{K}$ and countable. Let $\mathcal{N} := \bigcup_{Y\in\mathcal{D}} \mathcal{N}_Y$: $\mathcal{N}$ is $\mathbb{P}$-null as a countable union of $\mathbb{P}$-null sets.

Fixing $\omega \in \Omega\backslash\mathcal{N}$, we have $\forall Y \in \mathcal{D}, \quad \mathcal{E}_p(Y;\Theta(\omega)) \xrightarrow[p\to+\infty]{} \mathcal{E}(Y)$, thus point-wise convergence on $\mathcal{D}$ of the (now) deterministic function $\mathcal{E}_p(\cdot;\Theta(\omega))$ to $\mathcal{E}$. Now, a consequence of Proposition 2.1 is that the family of functions $(Y \mapsto \mathcal{E}_p(Y;\Theta'))_{\Theta'\in(\mathbb{S}^{d-1})^p}$ is equi-continuous on any compact (thus on $\mathcal{K}$). As a consequence, the point-wise convergence on $\mathcal{D}$ implies the uniform convergence of $\mathcal{E}_p(\cdot;\Theta(\omega))$ to $\mathcal{E}$ on $\overline{\mathcal{D}} = \mathcal{K}$ (a detailed presentation of this classic result can be found in [31], Proposition 3.2). This holds for any event $\omega \in \Omega\backslash\mathcal{N}$, with $\mathbb{P}(\Omega\backslash\mathcal{N}) = 1$, thus the uniform convergence is almost-sure. □

To complement this uniform almost-sure convergence, we prove a uniform Central Limit result on the error process $\sqrt{p}(\mathcal{E}_p - \mathcal{E})$ on a fixed compact set $\mathcal{K}$. This result provides insight on the law of the approximation error, uniformly with respect to the position $Y \in \mathcal{K}$.

Theorem 2.4. *Let $\mathcal{K} \subset \mathbb{R}^{n\times d}$ be a compact and non-empty set. On this domain, we have the following uniform Central Limit convergence in distribution of the approximation error of the random process $\mathcal{E}_p$:*

$$\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathcal{K})} G, \tag{2.16}$$

where the convergence is in law in the sense of $\ell^\infty(\mathcal{K})$, the space of bounded functions $z : \mathcal{K} \longrightarrow \mathbb{R}$ equipped with the uniform norm. The limit process G is the centred Gaussian process on $\mathcal{K}$ of covariance

$$\mathbb{C}(G)[Y,Y'] = \int_{\mathbb{S}^{d-1}} w_\theta(Y)w_\theta(Y')\mathrm{d}\varpi(\theta) - \mathcal{E}(Y)\mathcal{E}(Y').$$

This result implies the convergence in law of the uniform error

$$\sqrt{p}\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} \|G\|_{\ell^\infty(\mathcal{K})}. \tag{2.17}$$

We provide the proof of Theorem 2.4 in Section A.1, along with a brief presentation of the Donsker class arguments at hand.

Remark 2.5. *Our Central Limit result from Theorem 2.4 allows one to build (uniform) confidence intervals for the approximation $\mathcal{E}_p \approx \mathcal{E}$ on any compact, but is of limited practical interest due to the complexity of estimating the Gaussian process G. Nevertheless, such confidence intervals provide additional theoretical insight on the Monte-Carlo approximation of the discrete SW distance.*

Our result (2.16) complements a result by Xi and Niles-Weed [51], which shows the following distributional convergence of a related process which is a function of θ:

$$\mathbb{H}_n := \left\{\sqrt{n}\left(\mathrm{W}_q^q(P_\theta\#\hat{\mu}_n, P_\theta\#\hat{\nu}_n) - \mathrm{W}_q^q(P_\theta\#\mu, P_\theta\#\nu)\right), \theta \in \mathbb{S}^{d-1}\right\} \xrightarrow[n\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathbb{S}^{d-1})} \mathbb{H},$$

where μ, ν are compactly supported probability measures, and $\hat{\mu}_n, \hat{\nu}_n$ are discrete empirical versions supported on n samples of respective laws μ, ν, and $\mathbb{H}$ is a centred Gaussian process on $\mathbb{S}^{d-1}$.

The distributional convergence in (2.16) also complements a (quantified) convergence in probability by Xu and Huang [52]. For $q \geq 1$ and $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$, let $M_q(\mu) := (\int \|x\|^q \mathrm{d}\mu)^{1/q}$ and $L := q\mathrm{W}_q^{q-1}(\mu,\nu)(M_q(\mu)+M_q(\nu))$. In [52], they prove (Proposition 4) that for any $\varepsilon, \delta > 0$, if $p \geq \frac{2L^2}{(d-1)\varepsilon^2}\log(\frac{2}{\delta})$, then

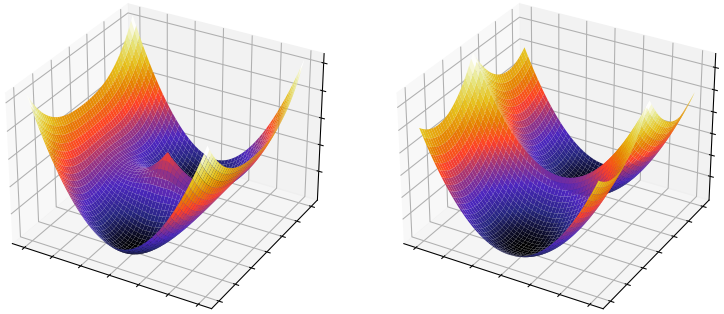
$$\mathbb{P}(|\widehat{\mathrm{SW}}_{q,p}^q(\mu,\nu) - \mathrm{SW}_q^q(\mu,\nu)| \geq \varepsilon) \leq \delta, \tag{2.18}$$

where $\widehat{\mathrm{SW}}_{q,p}^q(\mu,\nu)$ is the Monte-Carlo approximation of $\mathrm{SW}_q^q(\mu,\nu)$ with p projections. Their result is of a different nature, since it deals with general measures in $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ and does not study the process associated to moving the support of one of the measures. Point-wise, (2.18) from [52] is a more general result than (2.17), but the strength of our result comes from the study of the *process* $\mathcal{E}_p$, for which (2.18) is not informative in distribution, since it is a point-wise result. Furthermore, (2.18) is not tailored to our almost-sure uniform convergence case Theorem 2.3.

2.6. **Illustration in a simplified case.** Let us illustrate $\mathcal{E}$ in a simple case, that was briefly presented in Bonneel et al. [11], in order to grasp the difficulties at hand. This example is interesting for understanding the difficulty of performing computations with $\mathcal{E}$ and $\mathcal{E}_p$. Let $z_1 = (0,-1)^T$ and $z_2 = (0,1)^T$. Instead of computing $\mathcal{E}(Y)$ for any $Y \in \mathbb{R}^{2\times 2}$, we simplify by assuming $Y = (y,-y)^T = (y_1,y_2)^T$. We will assume further $y \neq 0$ and write $y = (u,v)^T$. The interested reader may seek the computations in Section A.2. With these notations, we can show that

$$\mathcal{E}(Y) = \mathrm{SW}_2^2(\gamma_Y,\gamma_Z) = \frac{u^2+v^2}{2} + \frac{1}{2} - \frac{2}{\pi}\left(|u| + |v|\,\mathrm{Arctan}\left|\frac{v}{u}\right|\right). \tag{2.19}$$

For W_2^2, one may show (see Section A.2 for the computations) that $\mathrm{W}_2^2(\gamma_Y,\gamma_Z) = u^2 + (|v|-1)^2$ in this setting. We compare $\mathcal{E}$ and W_2^2 in Figure 1.



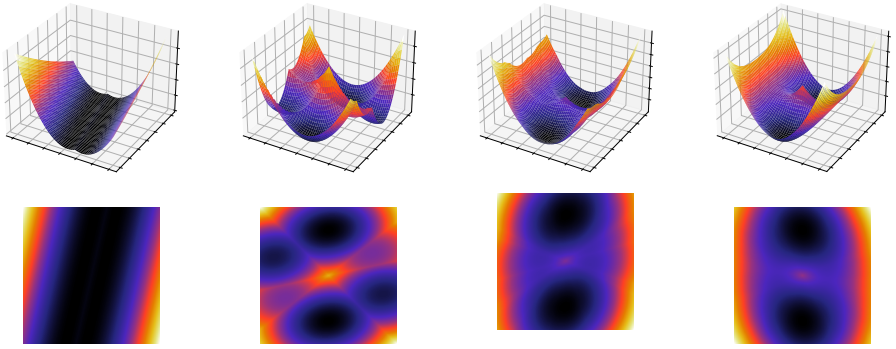
(A) $Y \longmapsto \mathcal{E}(Y)$ (B) $Y \longmapsto \mathrm{W}_2^2(\gamma_Y,\gamma_Z)$

FIGURE 1. Comparison between Sliced Wasserstein (a) and Wasserstein (b) landscapes for 2-point discrete measures $Y = (y,-y)^T$ and $Z = (z_1,z_2)^T$ with $z_1 = (0,-1)^T$ and $z_2 = (0,1)^T$.

Notice differences in regularity. $\mathcal{E}$ is smooth on the open set $\mathcal{U}$ (defined in (2.9)) of the $Y \in \mathbb{R}^{n\times d}$ with distinct points (this is known in general, [11]), but is not

differentiable anywhere in $\mathcal{U}^c$. Here this is clear at $(0,0)$. Furthermore, $\mathcal{E}$ presents two saddle points, $(\pm\frac{2}{\pi},0)$. In Section 3.1.2, we shall study the critical points of $\mathcal{E}$ in full generality. Finally, W_2^2 presents the typical landscape of the minimum of two quadratics.

We now move to computing $\mathcal{E}_p$ in this setting. In the case $n=2$, a significant simplification occurs since $\mathfrak{S}_2 = \{I, (2,1)\}$, and we express a simple formula for the cells in the Appendix, see Section A.2. We illustrate the cell structure in Figure 2.



(A) $\mathcal{E}_p$, $p=1$ (B) $\mathcal{E}_p$, $p=3$ (C) $\mathcal{E}_p$, $p=10$ (D) $\mathcal{E}$

FIGURE 2. The landscape $\mathcal{E}_p$ approaches $\mathcal{E}$ as p increases, but introduces numerous strict local optima. Notice that when p is too small ($p=1\leq d$ in particular), $\mathcal{E}_p$ even introduces other global optima.

Notice that as p increases, the number of new strict local optima also increases, however their associated cells become very small, thus one may hope that the probability of ending up in a strict local optimum would decrease as p increases. Specifically, in the heatmap visualisation, one may notice 6 large cells for $p=3$, and for $p=10$, two large cells corresponding to the global optima, and 8 small cells which may present local optima. This observation suggests that as $p\longrightarrow+\infty$, the total size of cells containing local optima decreases, and thus the probability of a numerical scheme converging to a local optimum decreases as well. Moreover, it is clear for the landscape $\mathcal{E}_p$ with $p=3$ that the critical points (points of differentiability with a null gradient) are exactly the minima of the cell quadratics. Remark that a cell may not contain the minimum of its quadratic, which is why we will refer to cells containing their minimum as "stable" (as is the case for all cells in $p=3$ illustration, but seemingly not for $p=10$).

As is suggested by Figure 2, even with a large number of projections p compared to the dimension d, the presence of strict local optima may prevent numerical solvers from converging to the global optimum $\gamma_Y=\gamma_Z$. This practical concern motivates the study of the landscapes $\mathcal{E}$ and $\mathcal{E}_p$, which is the topic of Section 3.

3. Properties of the Optimisation Landscapes of $\mathcal{E}$ and $\mathcal{E}_p$

The goal of this section is to study the respective landscapes of $\mathcal{E}$ and $\mathcal{E}_p$, their critical points and the links between them.

3.1. **Optimising $\mathcal{E}$.**

3.1.1. *Global optima of $\mathcal{E}$.* As its name suggests, the SW distance is indeed a distance on $\mathcal{P}_2(\mathbb{R}^d)$ (this result can be proven in the same manner for the q-SW distances, for $q \geq 1$).

Proposition 3.1 (Bonnotte [13], Theorem 5.1.2)**.** *SW is a distance on $\mathcal{P}_2(\mathbb{R}^d)$.*

As a consequence, the global optima of $\mathcal{E}$ are exactly the points Y^* such that $\gamma_{Y^*} = \gamma_Z$, or said otherwise the points such that $(y_1^*, \cdots, y_n^*)$ is a permutation of $(z_1, \cdots, z_n)$.

3.1.2. *Critical points of $\mathcal{E}$.* A first step in studying the landscape $\mathcal{E}$ is to determine its critical points, which we define as the set of points Y where $\mathcal{E}$ is differentiable and $\nabla\mathcal{E}(Y) = 0$. Thanks to Theorem 2.2, these critical points can be shown to satisfy a fixed point equation.

Corollary 3.1 (Equation characterising the critical points of $\mathcal{E}$)**.** *Let $Y \in \mathcal{U}$ (defined in (2.9)). For $(k,l) \in [\![1,n]\!]$, define $\Theta_{k,l}^{Y,Z} := \left\{\theta \in \mathbb{S}^{d-1} \mid \tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k) = l\right\} \subset \mathbb{S}^{d-1}$ and $S_{k,l}^{Y,Z} := d\int_{\Theta_{k,l}^{Y,Z}} \theta\theta^T \mathrm{d}\sigma \in S_d^+(\mathbb{R})$. Y is a critical point of $\mathcal{E}$ iif Y satisfies*

$$\forall k \in [\![1,n]\!],\ y_k = \sum_{l=1}^n S_{k,l}^{Y,Z} z_l. \tag{3.1}$$

Proof. Let $k \in [\![1,n]\!]$. We have $\mathbb{S}^{d-1} = \bigcup_{l=1}^n \Theta_{k,l}^{Y,Z}$, where the union is disjoint, therefore one may write

$$\begin{aligned}\frac{\partial\mathcal{E}}{\partial y_k}(Y) &= \frac{2}{n}\int_{\mathbb{S}^{d-1}} \theta\theta^T(y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)})\mathrm{d}\sigma(\theta) \\ &= \frac{2}{n}\sum_{l=1}^n \int_{\Theta_{k,l}} \theta\theta^T(y_k - z_l)\mathrm{d}\sigma(\theta) = \frac{2}{dn}y_k - \frac{2}{dn}\sum_{l=1}^n S_{k,l}^{Y,Z} z_l,\end{aligned}$$

where we have used $\int_{\mathbb{S}^{d-1}} \theta\theta^T \mathrm{d}\sigma(\theta) = I/d$ in the last equality. Equating the partial differential to 0 yields (3.1). □

Equation (3.1) shows that the critical points can be written as combinations of the points (z_l), "weighted" by the normalised conditional covariance matrices $S_{k,l}^{Y,Z} = d\mathbb{E}_{\theta\sim\sigma}\left[\mathbb{1}(\theta \in \Theta_{k,l}^{Y,Z})\theta\theta^T\right]$. With $\Psi := \begin{cases} \mathcal{U} & \longrightarrow & \mathbb{R}^{n\times d} \\ Y & \longmapsto & \begin{pmatrix} \sum_{l=1}^n z_l^T S_{1,l}^{Y,Z} \\ \vdots \\ \sum_{l=1}^n z_l^T S_{n,l}^{Y,Z} \end{pmatrix} \end{cases}$, Equation (3.1) writes as a fixed-point equation $Y = \Psi(Y)$.

Further notice that Ψ cannot be properly defined on $\mathcal{U}^c$, for instance if $n = 2$, and if $Y = (y, y)$, the two possible sorting choices $\tau_Y^\theta \in \{(1,2),(2,1)\}$ yield two different values for $\Psi(Y)$ (the first value is the second with the indices exchanged).

We show below that Ψ is continuous on $\mathcal{U}$. Unfortunately, Ψ cannot be extended to the whole space $\mathbb{R}^{n\times d}$, since the restrictions $\Psi|_{C_{\mathbf{m}}}$ may have distinct limits at the borders of the cells.

Proposition 3.2 (Regularity of Ψ)**.** *Ψ is continuous on $\mathcal{U}$ (defined in (2.9)).*

Proof. It is sufficient to prove the continuity of $G := Y \to S_{k,l}^{Y,Z}$ on $\mathcal{U}$, for k, l fixed. Let $Y \in \mathcal{U}$ and $\varepsilon > 0$. Define

(3.2)
$$\Theta_\varepsilon(Y) := \left\{ \theta \in \mathbb{S}^{d-1} \mid \forall \delta Y \in B(0,\varepsilon),\ \left(\theta^T y_{\tau_Y^\theta(k)} + \theta^T \delta y_{\tau_{\delta Y}^\theta(k)} \right)_{k\in[\![1,n]\!]} \in \mathcal{U}_{n,1} \right\},$$

with $\mathcal{U}_{n,1}$ the open set of lists $(x_1, \cdots, x_n) \in \mathbb{R}^n$ with distinct entries. By Bonneel et al. [11], Appendix A, Lemma 2, $\forall \theta \in \Theta_\varepsilon(Y),\ \forall \delta Y \in B(0,\varepsilon),\ \tau_Y^\theta = \tau_{Y+\delta Y}^\theta$. Let ε small enough such that $\forall \delta Y \in B(0,\varepsilon),\ Y + \delta Y \in \mathcal{U}$. Let $\delta Y \in B(0,\varepsilon)$. Separating the integral yields:

$$\begin{aligned} G(Y+\delta Y) &= \int_{\Theta_{k,l}^{Y+\delta Y,Z}} \theta\theta^T \mathrm{d}\varpi(\theta) \\ &= \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)} \theta\theta^T \mathrm{d}\varpi(\theta) + \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^T \mathrm{d}\varpi(\theta). \end{aligned}$$

Using the fact that $\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y) = \Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)$, and denoting $\|\cdot\|_{\mathrm{op}}$ the $\|\cdot\|_2$-induced operator norm on $\mathbb{R}^{d\times d}$, we get

$$\begin{aligned} G(Y+\delta Y) - G(Y) &= \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^T \mathrm{d}\varpi(\theta) - \int_{\Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^T \mathrm{d}\varpi(\theta), \\ \|G(Y+\delta Y) - G(Y)\|_{\mathrm{op}} &\le \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \left\|\theta\theta^T\right\|_{\mathrm{op}} \mathrm{d}\varpi + \int_{\Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \left\|\theta\theta^T\right\|_{\mathrm{op}} \mathrm{d}\varpi \\ &\le 2 \int_{\Theta_\varepsilon(Y)^c} 1 \mathrm{d}\varpi = 2\varpi(\Theta_\varepsilon(Y)^c). \end{aligned}$$

By Bonneel et al. [11], Appendix A, Lemma 3, there exists a constant C such that $\varpi(\Theta_\varepsilon(Y)^c) \le C\varepsilon$, which proves the continuity of G on $\mathcal{U}$. □

3.2. Optimising $\mathcal{E}_p$.

3.2.1. *Global optima of $\mathcal{E}_p$.* We saw in Proposition 3.1 that SW is a distance. Unfortunately, its discretised version $\widehat{\mathrm{SW}}_p$ is only a pseudo-distance: non-negativity, symmetry and the triangular inequality still hold, but separation fails.

For generic measures, a measure-theoretic way of seeing this is through characteristic functions. Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$, the condition $\widehat{\mathrm{SW}}_p(\mu,\nu) = 0$ is equivalent to $\forall i \in [\![1,p]\!],\ \forall t \in \mathbb{R},\ \phi_\mu(t\theta_i) = \phi_\nu(t\theta_i)$, where ϕ_μ (resp. ϕ_ν) is the characteristic function of μ (resp. ν). This condition only constrains the characteristic functions on p radial lines, and Bochner or Pólya-type criteria may be considered to find a characteristic function ϕ which equals ϕ_μ on these lines but differs on a non-null set.

The discrete case pertains more to our setting. As shown in [43], for p large enough, almost-sure separation holds. This result can be proven by leveraging the geometrical consequences of the constrains $P_{\theta_i}\#\gamma_Y = P_{\theta_i}\#\gamma_Z$, and determining the a.s. solution set using random affine geometry.

Theorem 3.1 ([43], Theorem 4)**.** *Let $\gamma_Z := \sum_{l=1}^{n} b_l \delta_{z_l}$, where the (z_l) are fixed and distinct. Assuming $\theta_1, \cdots, \theta_p \sim \mathfrak{o}^{\otimes p}$ and $n \geq 2$, we have*

- *if $p \leq d$, there exists $\mathfrak{o}$-a.s. an infinity of measures $\gamma \neq \gamma_Z \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\widehat{\mathrm{SW}}_p(\gamma, \gamma_Z) = 0$.*
- *if $p > d$, we have $\mathfrak{o}$-almost surely $\{\gamma_Z\} = \underset{\gamma \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \widehat{\mathrm{SW}}_p(\gamma, \gamma_Z)$.*

With a sufficient amount of projections, $\widehat{\mathrm{SW}}_p(\gamma_Y, \gamma_Z) = 0 \Rightarrow \gamma_Y = \gamma_Z$ (a.s.), hence when minimising $\widehat{\mathrm{SW}}_p(\gamma_Y, \gamma_Z)$ in Y, there is some hope of recovering γ_Z. Unfortunately, this does not guarantee that the (unique) solution will be attained numerically. This practical reality motivates the study of eventual local optima of $\mathcal{E}_p$.

The computation of the critical points of $\mathcal{E}_p$ can be done using the cell decomposition of Section 2.3. We show that the critical points of $\mathcal{E}_p$ are exactly the local optima of $\mathcal{E}_p$, and correspond to "stable cells", which is to say cells that contain the minimum of their quadratic.

3.2.2. *Critical points of $\mathcal{E}_p$ and cell stability.* The objective of this section is to confirm theoretically some of the intuitions provided by the illustrations of Section 2.6, namely that the critical points of $\mathcal{E}_p$ correspond to stable cells. Since the union of cells is exactly the differentiability set of $\mathcal{E}_p$, any critical point Y of $\mathcal{E}_p$ is necessarily within a cell $\mathcal{C}_{\mathbf{m}}$. Since $\mathcal{E}_p$ is quadratic on $\mathcal{C}_{\mathbf{m}}$, then a critical point Y is the minimum of the cell's quadratic $q_{\mathbf{m}}$. As a consequence, the critical points of $\mathcal{E}_p$ are exactly the "stable cell optima", i.e. the $Y \in \mathcal{U}$ (see the definition (2.9)) such that $Y = \underset{Y' \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} q_{\mathbf{m}(Y)}(Y')$.

The following theorem shows that there are no local optima of $\mathcal{E}_p$ outside of $\mathcal{U}$, and therefore that the set of local optima of $\mathcal{E}_p$, the set of critical points of $\mathcal{E}_p$ and the set of stable cell optima coincide. As previously, we define the set of critical points of $\mathcal{E}_p$ as the set of points Y where $\mathcal{E}_p$ is differentiable and $\nabla \mathcal{E}_p(Y) = 0$.

Theorem 3.2 (The local optima of $\mathcal{E}_p$ are within cells)**.** *Assume that $(\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p}$, then the following results hold $\mathfrak{o}$-almost surely. Let $Y \in \mathbb{R}^{n \times d}$ a local optimum of $\mathcal{E}_p$, then $\exists \mathbf{m} \in \mathfrak{S}_n^p$ such that $Y \in \mathcal{C}_{\mathbf{m}}$. As a consequence, we have the equality between the three sets:*

- *Local optima of $\mathcal{E}_p$;*
- *Critical points of $\mathcal{E}_p$;*
- *Stable cell optima: $\left\{ Y \in \mathcal{U} \mid Y = \underset{Y' \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} q_{\mathbf{m}(Y)}(Y') \right\}$.*

Proof. Let $Y \in \mathbb{R}^{n \times d}$ a local optimum of $\mathcal{E}_p$. Let $M := \{\mathbf{m} \in \mathfrak{S}_n^p \mid Y \in \overline{\mathcal{C}_{\mathbf{m}}}\}$.

Let $\mathbf{m} \in M$. Let us show that $\nabla q_{\mathbf{m}}(Y) = 0$ by contradiction: suppose $\nabla q_{\mathbf{m}}(Y) \neq 0$. For t positive and small enough,

$$\begin{aligned}\mathcal{E}_p(Y) &\leq \mathcal{E}_p\left(Y - t\frac{\nabla q_{\mathbf{m}}(Y)}{\|\nabla q_{\mathbf{m}}(Y)\|}\right) \leq q_{\mathbf{m}}\left(Y - t\frac{\nabla q_{\mathbf{m}}(Y)}{\|\nabla q_{\mathbf{m}}(Y)\|}\right) \\ &= q_{\mathbf{m}}(Y) - t\|\nabla q_{\mathbf{m}}(Y)\| + o(t) = \mathcal{E}_p(Y) - t\|\nabla q_{\mathbf{m}}(Y)\| + o(t).\end{aligned}$$

Therefore, for $t > 0$ sufficiently small, we have $\mathcal{E}_p(Y) < \mathcal{E}_p(Y)$, which is a contradiction. We now prove that $\#M = 1$. Using the notations of Remark 2.4, for $\mathbf{m} \in M$, we have $\nabla q_{\mathbf{m}}(Y) = 0$, thus $B\overrightarrow{y} = a_{\mathbf{m}}$. For $(\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p}$, we have $\mathfrak{o}$-almost surely that B is invertible and that $\mathbf{m} \neq \mathbf{m}' \implies a_{\mathbf{m}} \neq a_{\mathbf{m}'}$, thus $\mathfrak{o}$-almost surely, $\#M = 1$, proving that in fact Y belongs to $\mathcal{C}_{\mathbf{m}}$ and not to its boundary.

□

3.2.3. *Closeness of critical points of $\mathcal{E}_p$ and $\mathcal{E}$.* In practice, all numerical optimisation methods converge towards a local optimum. One may wonder what is the link between the critical points of $\mathcal{E}_p$, which we reach in practice, and the critical points of $\mathcal{E}$, among which are the theoretical solutions we would like to reach.

The following theorem shows that at the limit $p \to +\infty$, any sequence of critical points of $\mathcal{E}_p$ become fixed points of Ψ (3.1) in probability, which is to say that they exhibit similar properties to the critical points of $\mathcal{E}$.

Theorem 3.3 (Approximation of the fixed-point equation)**.**
For $p > d$, let Y_p any critical point of $\mathcal{E}_p$. Then we have the convergence in probability:

$$Y_p - \Psi(Y_p) \xrightarrow[p \longrightarrow +\infty]{\mathbb{P}} 0. \tag{3.3}$$

Specifically (see Corollary A.1), in order to reach a precision of ε, we have $\|Y_p - \Psi(Y_p)\|_{\infty,2} \leq \varepsilon$ with probability exceeding $1 - \eta$ if $p \geq \mathcal{O}\left(d^3 n \log(1/\eta)/\varepsilon^3\right)$ and $p \geq \mathcal{O}\left(d^3 n^2 \log(1/\eta)/\varepsilon^2\right)$, omitting logarithmic multiplicative terms in d and n.

We provide the proof in Section A.4, where we also estimate more precisely the convergence rate. The idea behind this result stems from computing the minima of the quadratics. Let $Y^* := \underset{Y}{\operatorname{argmin}}\, q_{\mathbf{m}}(Y)$, we have

$$y_k^* = A^{-1}\left(\frac{1}{p}\sum_{i=1}^{p} \theta_i \theta_i^T z_{\mathbf{m}_i(k)}\right) = \frac{A^{-1}}{p} \sum_{l \in [\![1,n]\!]} \sum_{\substack{i \in [\![1,p]\!] \\ \mathbf{m}_i(k) = l}} \theta_i \theta_i^T z_l, \tag{3.4}$$

with $A = \dfrac{1}{p}\displaystyle\sum_{i=1}^{p} \theta_i \theta_i^T$ which approaches the covariance matrix of $\theta \sim \mathfrak{o}$, i.e. I/d. Likewise, $\dfrac{1}{p} \displaystyle\sum_{\substack{i \in [\![1,p]\!] \\ \mathbf{m}_i(k) = l}} \theta_i \theta_i^T$ can be seen as an empirical conditional covariance, and it approaches $S_{k,l}^{Y,Z}/d$. We then apply matrix concentration inequalities to quantify the approximation error.

3.2.4. *Critical points of $\mathcal{E}_p$ and Block Coordinate Descent.* Leveraging on the cell structure of $\mathcal{E}_p$, we present an algorithm alternatively solving for the transport matrices and for the positions. Writing $\mathbb{U}$ the set of valid transport plans between two uniform measures with n points, we minimise the following energy (with $(\theta_1, \cdots, \theta_p)$ fixed)

$$J := \left\{ \begin{array}{ccc} \mathbb{U}^p \times \mathbb{R}^{n \times d} & \longrightarrow & \mathbb{R}_+ \\ (\pi^{(1)}, \cdots, \pi^{(p)}), Y & \longmapsto & \dfrac{1}{p}\displaystyle\sum_{i=1}^{p}\sum_{k=1}^{n}\sum_{l=1}^{n} (\theta_i^T y_k - \theta_i^T z_l)^2 \pi_{k,l}^{(i)} \end{array} \right. . \tag{3.5}$$

Observe that minimising J amounts to minimising $\mathcal{E}_p$.

Algorithm 1: Minimising $\mathcal{E}_p$ with Block-Coordinate Descent

Data: Fixed axes $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$, projections $(z_k^T \theta_i)_{k \in [\![1,n]\!],\, i \in [\![1,p]\!]}$.
Result: Positions $Y \in \mathbb{R}^{n \times d}$.
1 **Initialisation:** Draw $Y^{(0)} \in \mathbb{R}^{n \times d}$;
2 **for** $t \in [\![1, T_{\max}]\!]$ **do**
3 Update the OT maps by solving $\pi^{(t)} \in \underset{\pi \in \mathbb{U}^p}{\operatorname{argmin}}\ J(\pi, Y^{(t-1)})$;
4 Update the positions by solving $Y^{(t)} = \underset{Y \in \mathbb{R}^{n \times d}}{\operatorname{argmin}}\ J(\pi^{(t)}, Y)$;
5 **if** $\|Y^{(t)} - Y^{(t-1)}\|_{\infty,2} < \varepsilon$ **then**
6 Declare convergence and terminate.
7 **end**
8 **end**

The computation in Algorithm 1, line 3 is done using standard 1D OT solvers [22], and the update on the positions at line 4 can be computed in closed form (we provide the closed-form expression in Section A.5 for the sake of reproducibility). BCD can be seen as a walk from cell to cell (see Section 2.3), as illustrated in Figure 3. BCD moves from cell to cell and converges towards a stable cell optimum, and thus towards a local optimum of $\mathcal{E}_p$ (since these two sets are equal by Theorem 3.2). This behaviour is further studied in the experimental section.

4. Stochastic Gradient Descent on $\mathcal{E}$ and $\mathcal{E}_p$

As seen in Section 3.1.2, the optimisation properties of $\mathcal{E}$ and $\mathcal{E}_p$ indicate that optimising their landscapes might prove difficult in practice. In real-world applications, these landscapes (and especially $\mathcal{E}$, which is the most used) are minimised using Stochastic Gradient Descent. Perhaps unsurprisingly given the difficulties presented in Section 3.1.2 and due to the non-differentiable and non-convex properties of the landscapes, there has been no attempt to prove the convergence of such SGD schemes in the literature (to our knowledge). This section aims to bridge this knowledge gap, using recent theoretical results on the convergence of constant-step SGD schemes due to Bianchi et al. [6], and using results on decreasing-step SGD by Davis et al. [18]. Related works include Minibatch Wasserstein [21] in particular Section 5 wherein they leverage another non-convex non-differentiable SGD convergence framework from Majewski et al. [34] in order to derive convergence results for minibatch gradient descent on the Wasserstein and entropic Wasserstein distances.

There are other frameworks than Bianchi et al. [6] or Davis et al. [18] that one may consider in order to prove non-smooth, non-convex convergence of SGD, in particular the work of Majewski et al. [34]. Unfortunately, this work focuses on the case of *tame* functions, which is to say either *Clarke regular* functions ([14]), or *stratifiable functions* ([9]). It is not known whether $\mathcal{E}$ is Clarke regular (the graph in Figure 1 could intuitively point to the contrary, due to the local shape in $-\|x\|$, and it is known that $-\|\cdot\|$ is not Clarke regular). Likewise, it is not known whether $\mathcal{E}$ is stratifiable. Thankfully, our regularity results from Section 2 will allow us to

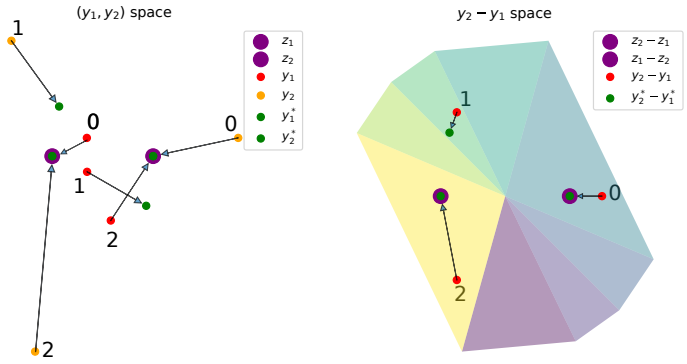


FIGURE 3. Illustration of the cell structure for $p = 4$ in dimension 2 from a BCD viewpoint. On the left, we view different points $Y = (y_1, y_2)$ (in red and orange) and the minima of their respective quadratics: (y_1^*, y_2^*), which should be compared to the original points (z_1, z_2) in purple. On the right, we view the cell structure depending on the position of $y_2 - y_1 \in \mathbb{R}^2$, since the cell conditions only depend on this difference (see (A.9)). We can see that in this example all cells are stable, thus there are three strict local optima of $\mathcal{E}_p$ in addition to the global optimum. The (y_1, y_2) pair number 0 is sent to (z_2, z_1), while the pair "1" is sent to a local optimum, and the pair "2" is sent to (z_1, z_2).

show that $\mathcal{E}$ is *path differentiable*, which is another (more general) regularity class which is enough to apply the results from Bianchi et al. [6] and Davis et al. [18].

Let us also mention the very recent work [32] (contemporary to ours), which studies the convergence of stochastic gradient schemes with decreasing steps and applied directly on probability measures instead of point clouds. Working with absolutely continuous measures μ and ν, the authors consider a scheme of the form

$$\mu^{(t+1)} = \left((1 - \alpha^{(t)})I + \alpha^{(t)} T_{\theta^{(t)}, \mu^{(t)}, \nu}\right) \# \mu^{(t)}, \tag{4.1}$$

where the $\theta^{(t)}$ are i.i.d. drawn from the uniform measure on the sphere, and where $T_{\theta^{(t)}, \mu^{(t)}, \nu}(x) := x + \left(\tau^{(t)}((\theta^{(t)})^T x) - (\theta^{(t)})^T x\right) \theta^{(t)}$, with $\tau^{(t)}$ the one-dimensional optimal transport map between the projected measures $P_{\theta^{(t)}} \# \mu^{(t)}$ and $P_{\theta^{(t)}} \# \nu$ (this map is uniquely defined on the support of $P_{\theta^{(t)}} \# \mu^{(t)}$ because $\mu^{(t)}$ and ν are absolutely continuous). It is quite easy to see that this scheme implements a stochastic gradient descent on $\mu \to \frac{1}{2}\mathrm{SW}_2^2(\mu, \nu)$. Building on proof techniques of [4], they show that if the sequence of learning rates $(\alpha^{(t)})_k$ is decreasing with $\sum \alpha^{(t)} = +\infty$ and $\sum (\alpha^{(k)})^2 < +\infty$, under some assumptions the sequence of measures $(\mu^{(t)})$ converges to ν for W_2. Unfortunately, extending these methods to discrete measures is not straightforward. Indeed, the authors of [4] claim that although they believe their results might hold for discrete measures, the difficulties of generalising their proofs to the discrete case were too important to achieve satisfying proofs in this case.

Before presenting our core results and the necessary theoretical framework from Bianchi et al. [6], we provide in Algorithm 2 the description of the SGD scheme used

to minimise either $\mathcal{E}$ or $\mathcal{E}_p$, i.e. for projections drawn with $\mu \in \{\varpi, \varpi_p\}$ respectively. Starting with random initial points $Y^{(0)} \sim \nu$, at each step t, we draw a random projection $\theta^{(t+1)} \sim \varpi$ and compute an SGD iteration of step $\alpha^{(t)}$ in the direction of the gradient of $Y \longmapsto w_{\theta^{(t+1)}}(Y)$. This scheme uses optionally an additive noise term controlled by a parameter a (that can be set to 0). In Section 4.2 to Section 4.3, we shall study constant-step SGD schemes, and in Section 4.5, we will focus on decreasing-step SGD.

Algorithm 2: Minimising $\mathcal{E}$ or $\mathcal{E}_p$ with Stochastic Gradient Descent

Data: Learning rate sequence $(\alpha^{(t)})_{t\in\mathbb{N}}$, noise level $a \geq 0$, convergence threshold $\beta > 0$, and probability distribution μ on $\mathbb{S}^{d-1}$.
Result: Positions $Y \in \mathbb{R}^{n\times d}$, assignment $\tau \in \mathfrak{S}_n$.

Initialisation: Draw $Y^{(0)} \in \mathbb{R}^{n\times d}$;
for $t \in [\![0, T_{\max} - 1]\!]$ **do**
  Draw $\theta^{(t+1)} \sim \mu$ and $\varepsilon^{(t+1)} \sim \mathcal{N}(0, I_{nd})$.
  SGD update:
  $$Y^{(t+1)} = Y^{(t)} - \alpha^{(t)} \left[\frac{\partial}{\partial Y} \mathrm{W}_2^2(P_{\theta^{(t+1)}} \# \gamma_Y, P_{\theta^{(t+1)}} \# \gamma_Z) \right]_{Y=Y^{(t)}} + \alpha^{(t)} a \varepsilon^{(t+1)}$$
  if $\|Y^{(t+1)} - Y^{(t)}\|_{\infty,2} < \beta$ **then**
    Declare convergence and terminate.
  end
end
return $Y^{(t_{\text{final}})}$ *and the assignment* τ *of* $\mathrm{W}_2^2(P_{\theta^{(t_{\text{final}})}} \# \gamma_{Y^{(t_{\text{final}})}}, P_{\theta^{(t_{\text{final}})}} \# \gamma_Z)$.

Overview of Main Results. In [6], Bianchi et al. establish conditions under which a constant-step SGD converges (in a certain sense), for a non-convex, locally Lipschitz cost function. Observe that both $\mathcal{E}$ and $\mathcal{E}_p$ are indeed locally Lipschitz, as shown in Theorem 2.1. In Section 4.2 and Section 4.3, we verify the required conditions for $\mathcal{E}$ and $\mathcal{E}_p$ (with p fixed projections), and prove results which can be broadly summarised as follows:

Theorem (Theorem 4.1: Convergence of the interpolated SGD (without noise) for $\mathcal{E}$ and $\mathcal{E}_p$)**.** *Given a sequence of SGD schemes $(Y_\alpha^{(t)})$ for $\mathcal{E}$ (resp. $\mathcal{E}_p$) of steps α, their associated piecewise affine interpolated schemes (Y_α) converge, in a weak sense as $\alpha \longrightarrow 0$, to the set of solutions of the differential inclusion equation $\dot{Y}(s) \in -\partial_C \mathcal{E}(Y(s))$ (resp. $\mathcal{E}_p$), where $\partial_C \mathcal{E}$ denotes the Clarke differential of $\mathcal{E}$.*

If we instead consider a *noised* SGD scheme (with noise magnitude $\alpha \times a,\ a > 0$), we have a stronger convergence result:

Theorem (Theorem 4.2: Convergence of the noised SGD for $\mathcal{E}$ and $\mathcal{E}_p$)**.** *Given a sequence of noised SGD schemes $(Y_\alpha^{(t)})$ for $\mathcal{E}$ (resp. $\mathcal{E}_p$) of steps α, they converge, in a weak sense as $\alpha \longrightarrow 0$, to the set of (Clarke) critical points of $\mathcal{E}$ (resp. $\mathcal{E}_p$).*

These results rely on the notion of Clarke differentiability, which generalises differentiability to non smooth functions as soon as these functions are locally Lipschitz (*i.e.* Lipschitz in a neighbourhood of each point). More precisely, for such a function $f : \mathbb{R}^d \to \mathbb{R}$, its Clarke sub-differential at x is defined as the convex hull of the limits of gradients of f

$$\partial_C f(x) := \mathrm{conv}\left\{ v \in \mathbb{R}^d : \exists (x_i) \in (\mathcal{D}_f)^{\mathbb{N}} : x_i \xrightarrow[i\longrightarrow+\infty]{} x \text{ and } \nabla f(x_i) \xrightarrow[i\longrightarrow+\infty]{} v \right\},$$

where $\mathcal{D}_f$ denotes the set of differentiability of f, whose complementary is of Lebesgue measure 0 by Rademacher's theorem, since f is locally Lipschitz. This notion of differentiability coincides with the classical one for differentiable functions, and with the usual sub-differential for convex functions. Clarke *critical points* of f are points x such that $0 \in \partial_C f(x)$.

In the case of decreasing-step SGD, in Section 4.5 we leverage the results of [18] to prove the following result, under typical assumptions on the decreasing steps $(\alpha^{(t)})$.

Theorem (Theorem 4.3: Convergence of decreasing-step noised SGD for $\mathcal{E}$ and $\mathcal{E}_p$)**.** *Consider $(Y^{(t)})$ a trajectory of noised decreasing-step SGD for $\mu \in \{\varpi, \varpi_p\}$ respectively, assume that it is almost-surely bounded. Then the sequence $F(Y^{(t)})$ is almost-surely convergent for $F \in \{\mathcal{E}, \mathcal{E}_p\}$ respectively, and almost-surely, any subsequential limit of $(Y^{(t)})$ belongs to the set of Clarke critical points of F.*

4.1. **Theoretical framework.** In the following, we briefly present the theoretical framework of Bianchi et al. [6]. They consider a function $f : \mathbb{R}^D \times \Theta \longrightarrow \mathbb{R}$, locally Lipschitz continuous in the first variable (for each θ), and μ a probability measure on $\Theta \subset \mathbb{R}^d$. Since f is locally Lipschitz in the first variable, the gradient $\nabla f(\cdot, \theta)$ of $f(\cdot, \theta)$ (w.r.t. the first variable) can be defined almost everywhere on $\mathbb{R}^D$, and any function $\varphi : \mathbb{R}^D \times \Theta \longrightarrow \mathbb{R}^D$ such that $\lambda \otimes \mu$ a.e., $\varphi = \nabla f$ is called an almost-everywhere gradient of f (see [6], Definition 1). Let $F := Y \longrightarrow \int_\Theta f(Y, \theta) \mathrm{d}\mu(\theta)$. A SGD scheme of step $\alpha > 0$ for F is a sequence $(Y^{(t)})$ of the form:

$$Y^{(t+1)} = Y^{(t)} - \alpha \varphi(Y^{(t)}, \theta^{(t+1)}), \quad \left(Y^{(0)}, (\theta^{(t)})_{t\in\mathbb{N}}\right) \sim \nu \otimes \mu^{\otimes \mathbb{N}}, \tag{4.2}$$

where ν is the distribution of the initial position $Y^{(0)}$, which we shall assume to be absolutely continuous w.r.t. the Lebesgue measure.

Within this framework, we can define an SGD scheme for $\mathcal{E}$ and $\mathcal{E}_p$. The function w_θ (Equation (2.5)) plays the role of f. We know from Proposition 2.1 that w_θ is locally Lipschitz (uniformly in θ), hence differentiable almost everywhere, and that at these points of differentiability, using [11] (Appendix A, "proof of differentiability"), the derivative of w_θ in Y is

$$\varphi(Y, \theta) := \left[\frac{2}{n} \theta\theta^T \left(y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)} \right) \right]_{k \in [\![1,n]\!]}, \tag{4.3}$$

which corresponds to the definition of an almost-everywhere gradient as proposed by [6]. Moreover, φ can be extended everywhere by choosing the sorting permutations arbitrarily when there is ambiguity. Within this framework, given a step $\alpha > 0$, and an initial position $Y^{(0)} \sim \nu$, the fixed-step SGD iterations (4.2) can be applied to $F = \mathcal{E}$ by choosing $\mu = \varpi$ or to $F = \mathcal{E}_p$ by choosing $\mu = \varpi_p :=$

$\frac{1}{p}\sum_i^p \delta_{\theta_i}$. We assume that $\mathrm{Span}(\theta_i)_{i\in[\![1,p]\!]} = \mathbb{R}^d$, which is satisfied $\mathfrak{o}$-almost surely if $(\theta_i)_{i\in[\![1,p]\!]} \sim \mathfrak{o}^{\otimes p}$, since $p > d$.

4.2. **Convergence of piecewise affine interpolated SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$.** The *piecewise-affine interpolated SGD scheme* associated to a discrete SGD scheme $(Y_\alpha^{(t)})$ of step α is defined as:

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

We consider the space of absolutely continuous curves from $\mathbb{R}_+$ to $\mathbb{R}^D$, denoted $\mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^D)$, and endow it with the metric of uniform convergence on all segments:

$$d_c(Y, Y') := \sum_{k\in\mathbb{N}^*} \frac{1}{2^k} \min\left(1, \max_{s\in[0,k]} \|Y(s) - Y'(s)\|_{\infty,2}\right). \tag{4.4}$$

We will show that when the step decreases, the interpolated processes approach the set of solutions of a differential inclusion equation. To that end, we define the set of absolutely continuous curves that start within a given compact $\mathcal{K}$ of $\mathbb{R}^D$ and are a.e. solutions of the differential inclusion:

$$S_{-\partial_C F}(\mathcal{K}) := \left\{Y \in \mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^D) \mid \underline{\forall} s \in \mathbb{R}_+,\ \dot{Y}(s) \in -\partial_C F(Y(s));\ Y(0) \in \mathcal{K}\right\}, \tag{4.5}$$

where $\underline{\forall}$ denotes "for almost every". Bianchi et al. [6] present three conditions under which they prove the convergence (in a certain weak sense) of interpolated SGD schemes on F. For the sake of self-containedness, we reproduce them here and verify them successively. Recall that for our two respective applications, $f(Y, \theta) = w_\theta(Y)$, $\mu \in \{\mathfrak{o},\ \mathfrak{o}_p\}$ and $F \in \{\mathcal{E},\ \mathcal{E}_p\}$.

Assumption 1.

i) *There exists $\kappa : \mathbb{R}^D \times \mathbb{S}^{d-1} \longrightarrow \mathbb{R}_+$ measurable such that each $\kappa(X, \cdot)$ is μ-integrable, and:*

$$\exists \varepsilon > 0,\ \forall Y, Y' \in B(X, \varepsilon),\ \forall \theta \in \mathbb{S}^{d-1},\ |f(Y, \theta) - f(Y', \theta)| \le \kappa(X, \theta)\|Y - Y'\|.$$

ii) *There exists $X \in \mathbb{R}^D$ such that $f(X, \cdot)$ is μ-integrable.*

Since f is the same in both cases, we can satisfy Assumption 1 for both schemes simultaneously. The (quantified) uniformly locally Lipschitz property of the w_θ (Proposition 2.1) allows us to verify Assumption 1, by letting $r := 1$ and $\kappa(X, \theta) := \kappa_1(X)$. Assumption 1 ii) is immediate since for *all* $Y \in \mathbb{R}^D$, $\theta \longmapsto w_\theta(Y)$ is continuous, therefore $\mathfrak{o} - L^1$ and $\mathfrak{o}_p - L^1$.

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

i) *There exists $c \ge 0$ such that $\forall X \in \mathbb{R}^D$, $\displaystyle\int_{\mathbb{S}^{d-1}} \kappa(X, \theta)\mathrm{d}\mu(\theta) \le c(1 + \|X\|)$.*

ii) *For every $\mathcal{K}$ compact of $\mathbb{R}^D$, $\displaystyle\sup_{X\in\mathcal{K}} \int_{\mathbb{S}^{d-1}} \kappa(X, \theta)^2 \mathrm{d}\mu(\theta) < +\infty$.*

The choice $\kappa(X, \theta) := \kappa_1(X)$ (independent on θ, and as defined in Proposition 2.1) satisfies Assumption 2. We now consider the Markov kernel associated to the SGD schemes, denoting the Borel sets $\mathcal{B}(\mathbb{R}^D)$:

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

With λ denoting the Lebesgue measure on $\mathbb{R}^D$, let

$$\Gamma := \{\alpha \in]0, +\infty[\mid \forall \rho \ll \lambda,\ \rho P_\alpha \ll \lambda\}.$$

We will verify the following assumption for both schemes:

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

In Proposition 4.1, we prove a stronger result, namely that Γ contains $]0, \frac{n}{2}[$, which allows us to simply take learning rates $0 < \alpha < \frac{n}{2}$, instead of having to specify $\alpha \in \Gamma$. As a weaker alternative, Assumption 3 could also be verified by noticing that for any $\theta \in \mathbb{S}^{d-1}$, the function w_θ is of class $\mathcal{C}^2$ almost everywhere (as detailed in Section 2.3), which allows us to apply [6] Proposition 4 and shows that Assumption 3 holds.

Proposition 4.1. *For schemes (4.2) applied to* $\mathcal{E}$ *or* $\mathcal{E}_p$, $\Gamma = \mathbb{R}_+^* \setminus \{\frac{n}{2}\}$.

Proof. Let $\mu \in \{\varpi,\ \varpi_p\}$. Recall the line-by-line notation $Y = (y_1, \cdots, y_n)^T \in \mathbb{R}^{n\times d}$. We also denote $Z_\tau := (z_{\tau(1)}, \cdots, z_{\tau(n)})^T$ for $\tau \in \mathfrak{S}_n$. Let $\rho \ll \lambda$ and $B \in \mathcal{B}(\mathbb{R}^D)$ such that $\lambda(B) = 0$. We have, with $\alpha' := 2\alpha/n$:

$$\begin{aligned}
\rho P_\alpha(B) &= \int_{\mathbb{R}^D}\int_{\mathbb{S}^{d-1}} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^T) + \alpha' Z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}}\theta\theta^T\right)\mathrm{d}\mu(\theta)\mathrm{d}\rho(Y)\\
&\le \sum_{\tau\in\mathfrak{S}_n}\int_{\mathbb{R}^D}\int_{\mathbb{S}^{d-1}} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^T) + \alpha' Z_\tau\theta\theta^T\right)\mathrm{d}\mu(\theta)\mathrm{d}\rho(Y)\\
&= \sum_{\tau\in\mathfrak{S}_n}\int_{\mathbb{S}^{d-1}} I_\tau(\theta)\mathrm{d}\mu(\theta),
\end{aligned}$$

where $I_\tau(\theta) := \displaystyle\int_{\mathbb{R}^D} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^T) + \alpha' Z_\tau\theta\theta^T\right)\mathrm{d}\rho(Y)$, and where the last line is obtained by applying Tonelli's theorem. Let $\tau \in \mathfrak{S}_n$ and $\theta \in \mathbb{S}^{d-1}$. We now assume $\alpha' \neq 1$, which is to say $\alpha \neq n/2$. We operate the affine change of variables $X = \phi(Y) := Y(I-\alpha'\theta\theta^T) + \alpha' Z_\tau\theta\theta^T$, which is invertible for $\alpha' \neq 1$. We have

$$I_\tau(\theta) = \int_{\mathbb{R}^D} \mathbb{1}_B(\phi(Y))\mathrm{d}\rho(Y) = \int_{\mathbb{R}^D} \mathbb{1}_B(X)\mathrm{d}\phi\#\rho(X) = \phi\#\rho(B).$$

Now since ϕ is affine and invertible, $\phi\#\rho \ll \lambda$, thus $\phi\#\rho(B) = 0$, and finally $\rho P_\alpha(B) = 0$. This proves that $\rho P_\alpha \ll \lambda$ for $\alpha > 0$ differing from $n/2$. □

Now that we have verified Assumption 1, Assumption 2 and Assumption 3, we can apply [6], Theorem 2 to $\mathcal{E}$ and $\mathcal{E}_p$. Let $0 < \alpha_0 < n/2$.

Theorem 4.1 ([6], Theorem 2 applied to $\mathcal{E}$ and $\mathcal{E}_p$: convergence of the interpolated SGD scheme)**.** *Let* $(Y_\alpha^{(t)}), \alpha \in]0, \alpha_0], t \in \mathbb{N}$ *a collection of SGD sequences associated to (4.2) applied to* $\mathcal{E}$ *or* $\mathcal{E}_p$*. Consider* (Y_α) *their associated piecewise affine interpolations. For any* $\mathcal{K}$ *compact of* $\mathbb{R}^D$ *and any* $\varepsilon > 0$*, we have for* $F \in \{\mathcal{E},\ \mathcal{E}_p\}$ *and* $\mu \in \{\varpi,\ \varpi_p\}$ *respectively*

$$\lim_{\substack{\alpha \xrightarrow{} 0\\ \alpha \in\ ]0,\alpha_0]}} \nu \otimes \mu^{\otimes\mathbb{N}}\left(d_c(Y_\alpha, S_{-\partial_C F}(\mathcal{K})) > \varepsilon\right) = 0, \tag{4.6}$$

where d_c *is the metric of uniform convergence defined in (4.4).*

It is to be understood that when the SGD step decreases, the interpolated schemes converge towards the set of solutions of the differential inclusion related to the continuous SGD equation. This convergence is weak: the distance to this set approaches 0 in probability, and $S_{-\partial_C F}(\mathcal{K})$ is a set of solutions which we do not know how to compute, however we can study some theoretical properties of the solutions given a suitable starting point $Y(0)$, see Remark 4.1.

Remark 4.1. *For $\mathcal{E}$, if the initial position $Y^{(0)}$ belongs to a maximal connected component $\mathcal{V}$ of the differentiability set $\mathcal{U}$ (which is open), then consider the gradient flow differential equation*

$$\frac{\partial \gamma}{\partial t}(Y,s) = -\nabla \mathcal{E}(\gamma(Y,s)), \quad \gamma(Y,0) = Y, \quad \gamma(Y,s) \in \mathcal{V}. \tag{4.7}$$

Since $\mathcal{E}$ is of class $\mathcal{C}^1$ on $\mathcal{V}$ (by Theorem 2.2), with $\nabla\mathcal{E}$ Lipschitz (locally would suffice), standard flow results show that there exists a unique solution $\gamma(Y,\cdot)$ for any $Y \in \mathcal{V}$ defined on some interval $]a_Y, b_Y[\subset \mathbb{R}$, which defines a continuous function $\gamma : \mathcal{D} \longrightarrow \mathcal{V}$, with $\mathcal{D} = \{(s,Y) \in \mathbb{R} \times \mathcal{V} \mid s \in]a_Y, b_Y[\}$. Since in our case, we consider a gradient flow, and since for any $c \in \mathbb{R}$, the set $A_c := \{Y \in \mathcal{V} \mid \mathcal{E}(Y) \leq c\}$ is compact, in fact the flows $\gamma(Y,s)$ are defined for $s \in [0, +\infty[$. Furthermore, if a sequence $(\gamma(Y,s_m))_{m\in\mathbb{N}}$ were to converge to a limit Y^∞, then one would have $Y^\infty \in \mathcal{V}$ and $\nabla\mathcal{E}(Y^\infty) = 0$. Our work does not show that the set $\mathcal{Z}_\mathcal{V} := \{Y \in \mathcal{V} \mid \nabla\mathcal{E}(Y) = 0\}$ of critical points of $\mathcal{E}$ is finite, however if that were the case, then more standard euclidean gradient flow results show that for any $Y \in \mathcal{V}$, $\exists Y^\infty \in \mathcal{Z}_\mathcal{V} :\ \gamma(Y,s) \xrightarrow[s\longrightarrow+\infty]{} Y^\infty$.

Note that given a learning rate $\alpha > 0$, an SGD scheme (4.2) applied to $\mathcal{E}$ and starting in $\mathcal{V}$ has no reason to stay in $\mathcal{V}$, and we unfortunately do not have equality with a discretised version of the gradient flow. However, thanks to Bianchi et al. [6] *Theorem 1, the trajectories stay almost-surely in differentiability points of $\mathcal{E}$ and w_θ, and thus almost-surely, $\varphi(Y^{(t)}, \theta^{(t+1)}) = \nabla w_{\theta^{(t+1)}}(Y^{(t)})$.*

4.3. **Convergence of Noised SGD Schemes on $\mathcal{E}$ and $\mathcal{E}_p$.** In order to prove stronger convergence results we need to consider noised variants of our SGD schemes. Consider $\varepsilon \sim \eta := \mathcal{N}(0, I_D)$ an independent noise, our schemes become:

$$Y^{(t+1)} = Y^{(t)} - \alpha\varphi(Y^{(t)}, \theta^{(t+1)}) + \alpha a \varepsilon^{(t+1)}, \quad (Y^{(0)}, (\theta^{(t)})_{t\in\mathbb{N}}, (\varepsilon^{(t)})_{t\in\mathbb{N}}) \sim \nu \otimes \mu^{\otimes\mathbb{N}} \otimes \eta^{\otimes\mathbb{N}} \tag{4.8}$$

where $\mu = \varpi$ for $\mathcal{E}$ and ϖ_p for $\mathcal{E}_p$. We follow the method from [6], which suggests that adding a small perturbation (that decreases with the step size) allows us to verify additional suitable assumptions. Note that this modification does not impact our verification of the previous assumptions 1 through 3. Bianchi et al. introduce the following assumption:

Assumption 4. *there exists $V, p : \mathbb{R}^D \to \mathbb{R}_+$ and $\beta : \mathbb{R}_+^* \to \mathbb{R}_+^*$ measurable, as well as $C \geq 0$, such that for any $\alpha \in \Gamma \cap]0, \alpha_0]$:*

i) *$\exists R(\alpha) > 0,\ \delta(\alpha) > 0,\ \exists \rho(\alpha)$ a probability measure on $\mathbb{R}^D$, such that:*

$$\forall Y \in \overline{B}(0,R),\ \forall A \in \mathcal{B}(\mathbb{R}^D),\ P_\alpha(Y,A) \geq \delta\rho(A).$$

ii) $\sup_{Y\in\overline{B}(0,R)} V(Y) < +\infty$ *and* $\inf_{Y\in\overline{B}(0,R)^c} p(Y) > 0$, *with:*

$$\forall Y \in \mathbb{R}^D,\ P_\alpha V(Y) \leq V(Y) - \beta(\alpha)p(Y) + C\beta(\alpha)\mathbb{1}_{\overline{B}(0,R)}(Y).$$

iii) $p(Y) \xrightarrow[Y \longrightarrow \infty]{} +\infty$.

Thanks to Bianchi et al. [6], Proposition 5, this noised setting implies immediately Assumption 4 i), for *any* choice of $R > 0$. They also suggest more restrictions on f that imply Assumption 4 ii) and iii), which our use case does not satisfy. We shall verify Assumption 4 ii) and iii) for $\mathcal{E}$ and $\mathcal{E}_p$ separately, but using similar methods. Beforehand, let us remark that the Markov kernel associated to (4.8) is determined by the following action on measurable functions $g : \mathbb{R}^D \longrightarrow \mathbb{R}$:

$$P_\alpha g(Y) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}^D} g(Y - \alpha\varphi(Y, \theta) + \alpha a X) \mathrm{d}\mu(\theta)\mathrm{d}\eta(X).$$

Proposition 4.2 (Drift property for noised SGD on $\mathcal{E}$)**.** *Let $V := \|\cdot\|_F^2$, $\alpha_0 < 1$ and $p(Y) := \dfrac{2}{dn}(1-\alpha_0)\displaystyle\sum_{k=1}^{n} \|y_k\|_2^2$. There exists $R > 0$ and $C \geq 0$:*

$$\forall \alpha \in \,]0, \alpha_0], \ \forall Y \in \mathbb{R}^D, \ P_\alpha V(Y) \leq V(Y) - \alpha p(Y) + C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y).$$

Therefore, Assumption 4 is satisfied for (4.8) when $\mu = \varpi$.

Proof. Let $Y \in \mathbb{R}^D$, $\alpha \in]0, 1[$ and $\Delta(Y) := P_\alpha V(Y) - V(Y)$. We expand the square, then leverage the fact that η is centred, and decompose, writing $\varphi_k := \varphi(Y, \theta)_{k,\cdot}$:

$$\Delta(Y) = \alpha^2 a^2 nd + \underbrace{\alpha^2 \sum_{k=1}^{n} \int_{\mathbb{S}^{d-1}} \varphi_k^T \varphi_k \mathrm{d}\varpi(\theta)}_{\Delta_1(Y)} \underbrace{-2\alpha \sum_{k=1}^{n} \int_{\mathbb{S}^{d-1}} y_k^T \varphi_k \mathrm{d}\varpi(\theta)}_{\Delta_2(Y)}.$$

We have $\varphi_k^T \varphi_k = \dfrac{4}{n^2}(y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)})^T \theta\theta^T (y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)})$. Then recall that for all $\theta \in \Theta_{k,l}^{Y,Z}$, $z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)} = z_l$. It follows that

$$\begin{aligned}
\Delta_1(Y) &= \frac{4\alpha^2}{n^2} \sum_{(k,l) \in [\![1,n]\!]^2} \int_{\Theta_{k,l}^{Y,Z}} (x_k - z_l)^T \theta\theta^T (y_k - z_l) \mathrm{d}\varpi(\theta) \\
&= \frac{4\alpha^2}{dn^2} \sum_{(k,l) \in [\![1,n]\!]^2} (y_k - z_l)^T S_{k,l}^{Y,Z} (y_k - z_l) \\
&\leq \frac{4\alpha\alpha_0}{dn^2} \sum_{(k,l) \in [\![1,n]\!]^2} \|y_k - z_l\|_2^2 \leq \frac{4\alpha\alpha_0}{dn} \left(\sum_{k=1}^{n} (\|y_k\|_2^2 + 2\|Z\|_{\infty,2} \|y_k\|_2) + n\|Z\|_{\infty,2}^2 \right),
\end{aligned}$$

where we used the inequality $S_{k,l}^{Y,Z} \preceq I_d$.

Now for Δ_2, we have $y_k^T \varphi_k = \frac{2}{n}(\theta^T y_k)\theta^T (y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)})$, hence

$$\begin{aligned}
\Delta_2(Y) &= -\frac{4\alpha}{dn} \sum_{(k,l) \in [\![1,n]\!]^2} y_k^T S_{k,l}^{Y,Z} (y_k - z_l) \\
&= -\frac{4\alpha}{dn} \sum_{k=1}^{n} \|y_2\|^2 + \frac{4\alpha}{dn} \sum_{(k,l) \in [\![1,n]\!]^2} y_k^T S_{k,l}^{Y,Z} z_l \\
&\leq -\frac{4\alpha}{dn} \sum_{k=1}^{n} \|y_2\|^2 + \frac{4\alpha}{d} \|Z\|_{\infty,2} \sum_{k=1}^{n} \|y_k\|_2,
\end{aligned}$$

since $\sum_{l=1}^{n} S_{k,l}^{Y,Z} = I_d$ and $S_{k,l}^{Y,Z} \preceq I_d$. Finally,

$$\Delta(Y) \leq \alpha \Bigg[\underbrace{-\frac{4}{dn}(1-\alpha_0)\sum_{k=1}^{n}\|y_k\|_2^2}_{q(Y)} + \underbrace{\alpha_0 a^2 nd + \frac{4\alpha_0}{d}\|Z\|_{\infty,2}^2 + \frac{4}{d}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2 + \frac{8\alpha_0}{dn}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2}_{r(Y)} \Bigg].$$

Now since $\frac{r(Y)}{q(Y)} \xrightarrow[\|Y\|\longrightarrow+\infty]{} 0$, there exists $R > 0$ such that for $Y \in \mathbb{R}^D$ such that $\|Y\|_{\infty,2} > R$, we have $r(Y) \leq q(Y)/2$. In that case, we have $\Delta(Y) \leq \alpha(-q(Y) + q(Y)/2) = -\alpha q(Y)/2$. For $Y \in \mathbb{R}^D$ such that $\|Y\|_{\infty,2} \leq R$, we have $\Delta(Y) \leq \alpha r(Y) \leq \alpha \max_{\|Y\|_{\infty,2}\leq R} r(Y) =: C\alpha$ (C exists since r is continuous on the compact $\overline{B}(0,R)$.) This proves that for any $Y \in \mathbb{R}^D$, $\Delta(Y) \leq -\alpha q(Y)/2 + C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y)$. □

We now turn to the scheme for $\mathcal{E}_p$. Let $A := \frac{1}{p}\sum_{j=1}^{p}\theta_j\theta_j^T$, and consider $\lambda_{\min}(A)$ its smallest eigenvalue. Note that $\lambda_{\min}(A) > 0$, since we assumed $\mathrm{Span}(\theta_j)_{j\in[\![1,p]\!]} = \mathbb{R}^d$.

Proposition 4.3 (Drift property for noised SGD on $\mathcal{E}_p$)**.** *Let $V := \|\cdot\|_F^2$, $\alpha_0 < n$ and $q(Y) := \frac{2}{n}\left(1 - \frac{\alpha_0}{n}\right)\lambda_{\min}(A)\sum_{k=1}^{n}\|y_k\|_2^2$. There exists $R > 0$ and $C \geq 0$:*

$$\forall \alpha \in]0, \alpha_0],\ \forall Y \in \mathbb{R}^D,\ P_\alpha V(Y) \leq V(Y) - \alpha q(Y) + C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y).$$

Therefore, Assumption 4 is satisfied for (4.8) when $\mu = \varpi_p$.

We leverage the same strategy as Proposition 4.2, yet the technicalities of the upper-bounds differ.

Proof. Let $Y \in \mathbb{R}^D$ and $\alpha \in]0, \alpha_0]$. We expand the squares and use that η is centred:

$$\Delta(Y) := P_\alpha V(Y) - V(Y) = \alpha^2 a^2 nd + \underbrace{\alpha^2 \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{n}\sum_{i=1}^{d}\varphi(Y,\theta_j)_{k,i}^2}_{\Delta_1(Y)} \underbrace{-2\alpha\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{n}\sum_{i=1}^{d} y_{k,i}\varphi(Y,\theta_j)_{k,i}}_{\Delta_2(Y)}.$$

On the one hand,

$$\Delta_1(Y) = \frac{4\alpha^2}{pn^2}\sum_{j=1}^{p}\sum_{k=1}^{n}(y_k - z_{\tau_Z^{\theta_j}\circ(\tau_Y^{\theta_j})^{-1}(k)})^T\theta_j\theta_j^T(y_k - z_{\tau_Z^{\theta_j}\circ(\tau_Y^{\theta_j})^{-1}(k)}) \leq \frac{4\alpha\alpha_0}{n^2}\left(n\|Z\|_{\infty,2}^2 + \sum_{k=1}^{n}\left(y_k^T A y_k + 2\|Z\|_{\infty,2}\|y_k\|_2\right)\right).$$

Similarly, $\Delta_2(Y) \leq -\frac{4\alpha}{n}\sum_{k=1}^{n} y_k^T A y_k + \frac{4\alpha}{n}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2$. Let

$$q_0(Y) := \frac{4}{n}\left(1-\frac{\alpha_0}{n}\right)\lambda_{\min}(A)\sum_{k=1}^{n}\|y_k\|_2^2,$$

$$r(Y) := \alpha_0 a^2 n d + \frac{4\alpha_0}{n}\|Z\|_{\infty,2}^2 + \left(\frac{8\alpha_0}{n^2}+\frac{4}{n}\right)\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2.$$

We have $\Delta(Y) \leq \alpha(-q_0(Y)+r(Y))$, and we can conclude using the same method as Proposition 4.2. □

Finally, we require the fairly natural assumption that F admits a "chain rule".

Assumption 5.
For any $Y \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^D)$, $\underline{\forall} s > 0$, $\forall V \in \partial_C F(Y(s))$, $V^T \dot{Y}(s) = (F\circ Y)'(s)$.

In order to satisfy Assumption 5, we will use the following result:

Proposition 4.4. *Any* $F : \mathbb{R}^D \longrightarrow \mathbb{R}$ *locally Lipschitz and semi-concave admits a chain rule for the Clarke sub-differential, and thus satisfies Assumption 5.*

Proof. Let $F : \mathbb{R}^D \longrightarrow \mathbb{R}$ locally Lipschitz and semi-concave. By Vial (1983) [47], Proposition 4.5, this implies that $-F$ is Clarke regular. Then, by Bolte and Pauwels [10], Proposition 2, the fact that $-F$ is Clarke regular implies that F is path differentiable, and thus admits a chain rule, by Bolte et al. [10], Corollary 2. □

Since $\mathcal{E}$ is semi-concave (Proposition 2.4) and locally Lipschitz, Proposition 4.4 allows us to verify Assumption 5 for (4.8). We may follow the same line of thought for $\mathcal{E}_p$, or alternatively we may use the fact that it is semi-algebraic (Proposition 2.5). By Bolte and Pauwels (2021), [10], Proposition 2, this implies that $\mathcal{E}_p$ is path differentiable. Then by Bolte and Pauwels [10], Corollary 2, path differentiability implies having a chain rule for the Clarke sub-differential, which is verbatim [6], Assumption 5. We now have all the assumptions for [6], Theorem 3:

Theorem 4.2 (Applying [6], Theorem 3: convergence of noised SGD schemes to a critical point)**.** *Consider a collection of noised SGD schemes* $(Y_\alpha^{(t)})$*, associated to (4.8), respectively for* $F \in \{\mathcal{E},\ \mathcal{E}_p\}$*, with steps* $\alpha \in]0, \alpha_0]$*, with* $\alpha_0 < 1$*. Let* $\mathcal{Z}$ *the set of Clarke critical points of* F*, i.e.* $\mathcal{Z} := \left\{Y \in \mathbb{R}^D \mid 0 \in \partial_C F(Y)\right\}$*. For* $\mu \in \{\varpi,\ \varpi_p\}$ *respectively, we have:*

$$\forall \varepsilon > 0,\ \varlimsup_{t\longrightarrow+\infty}\ \nu\otimes\mu^{\otimes\mathbb{N}}\otimes\eta^{\otimes\mathbb{N}}\left(d(Y_\alpha^{(t)},\mathcal{Z}) > \varepsilon\right) \xrightarrow[\substack{\alpha\longrightarrow 0\\ \alpha\in]0,\alpha_0]}]{} 0.$$

It is to be understood that the euclidean distance between any sub-sequential limit of $(Y_\alpha^{(t)})_t$ and set of Clarke critical points $\mathcal{Z}$ approaches 0 in probability as the step size decreases. The distance d in the Theorem refers to the $\|\cdot\|_2$-induced distance between the point $Y_\alpha^{(t)} \in \mathbb{R}^D$ and the set $\mathcal{Z} \subset \mathbb{R}^D$.

Computing the set of Clarke critical points of $\mathcal{E}$ remains an open problem, and seems out of reach considering the difficulty of the simpler problem of computing the points where $\mathcal{E}$ is differentiable and $\nabla\mathcal{E} = 0$ (see the discussion in Section 3.1.2). The difficulty lies at the boundaries of $\mathcal{U}$ (see (2.9)), where there is no longer unicity of the sorting permutations of $(\theta^T x_k)_{k=1,\dots,n}$ and $(\theta^T z_l)_{l=1,\dots,n}$ for ϖ-almost-every

$\theta \in \mathbb{S}^{d-1}$. Computing the Clarke sub-differential at such points in closed form and determining the associated critical points seems out of reach since there is already no closed form for smooth critical points Corollary 3.1.

For $\mathcal{E}_p$, the set of Clarke critical points strictly contains the set of critical points established in Theorem 3.2. In general, the set of Clarke critical points that lie outside of the set of differentiability $\widetilde{\mathcal{Z}} := \mathcal{Z} \cap (\cup_{\mathbf{m}} \mathcal{C}_{\mathbf{m}})^c$ is not empty, yet by Theorem 3.2 it cannot contain a local optimum, and thus only contains saddle points. We believe that these saddle points will in practice never be the limit of our noised SGD trajectories, since intuition suggests that a trajectory attaining such a point at a certain time will find a decreasing direction almost-surely. Showing such a result rigorously is out of the scope of this paper since this question is still an active field in simpler smooth cases [27, 28]. More precisely, the minimisation of generic non-smooth non-convex functions F is also still actively studied. For instance, [16] and [7] investigate conditions to avoid convergence to certain saddle points, for randomly initialised deterministic (and potentially noised) proximal methods for semi-convex functions under novel strict saddle conditions. Another related reference is [17], which studies the non-convergence of noised sub-gradient descent to saddle points. We illustrate in Figure 4 the Clarke critical points of $\mathcal{E}_p$ for $p = 3$, on the numerical example of Section 2.6.

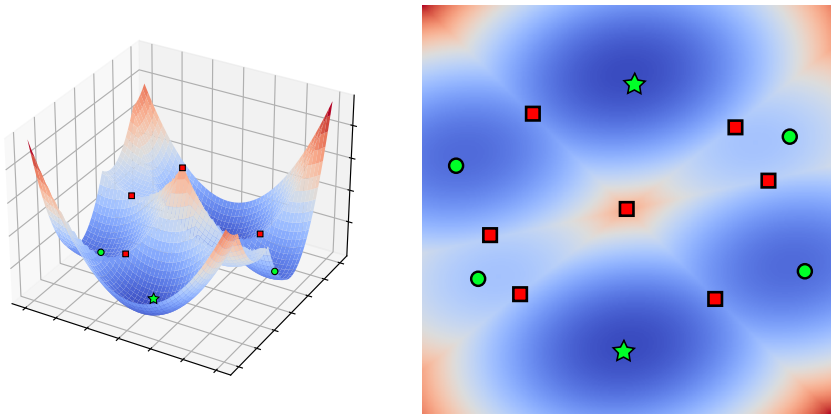


FIGURE 4. The stars, circles and squares are the Clarke critical points of $x \mapsto \mathcal{E}_p(X = (-x, x)^T)$, $x \in \mathbb{R}^2$ for $p = 3$. The squares do not correspond to local optima of $\mathcal{E}_p$, and are unlikely to be reached numerically. The circles and stars correspond to local optima of $\mathcal{E}_p$: the stars correspond to the global optima and satisfy the desired results $\mathcal{E}_p = 0$, while the circles are strict local optima.

In full generality (without the symmetry restriction, and with larger parameters p, n, d), one may expect that the Clarke critical points will have a similar structure.

4.4. **Discussion on result generalisation. Batching.** One may consider a variant in which at each step t, one draws a random batch of b directions independently from a measure μ over $\mathbb{S}^{d-1}$ ($\mu \in \{\sigma, \sigma_p\}$, for our purposes). Algorithmically, one

does the following SGD scheme:

$$Y^{(t+1)} = Y^{(t)} - \frac{\alpha}{b}\sum_{j=1}^{b}\varphi(Y^{(t)},\theta_j^{(t+1)}), \quad (Y^{(0)},(\theta_j^{(t)})_{\substack{j\in[\![1,b]\!]\\ t\in\mathbb{N}}}) \sim \nu\otimes(\mu^{\otimes b})^{\otimes\mathbb{N}}. \tag{4.9}$$

In order to fit our theoretical framework (see Section 4.1), we define

$$g(Y,(\theta_1,\cdots,\theta_b)) := \frac{1}{b}\sum_{j=1}^{b} f(Y,\theta_j).$$

Furthermore, the a.e. gradient of g becomes $\psi(\cdot,(\theta_1,\cdots,\theta_b)) := \frac{1}{b}\sum_{j=1}^{b}\varphi(\cdot,\theta_j)$ instead of $\varphi(\cdot,\theta^{(t)})$. The function over which (4.9) performs SGD is:

$$\begin{aligned} G(Y) &= \int_{\mathbb{S}^{d-1}} g(Y,\theta_1,\cdots,\theta_b)\mathrm{d}\mu^{\otimes b}(\theta_1,\cdots,\theta_p)\\ &= \int_{\mathbb{S}^{d-1}} \frac{1}{b}\sum_{j=1}^{b} f(Y,\theta_j)\mathrm{d}\mu^{\otimes b}(\theta_1,\cdots,\theta_p)\\ &= \int_{\mathbb{S}^{d-1}} f(Y,\theta)\mathrm{d}\mu(\theta) = F(Y). \end{aligned}$$

One may check easily that if Assumptions 1 through 5 of Section 4.1 are satisfied for (f,F), then they are satisfied for (g,G). As a consequence, all our results can be adapted without any difficulty to the batched setting.

Barycentres. If one were to replace $\mathcal{E}$ with the barycentre energy $\mathcal{E}_{\mathrm{bar}}$ (2.3), the sample loss would become

$$g(Y,\theta) = \sum_{j=1}^{J}\lambda_j f_j(Y,\theta_j), \text{ where } f_j(Y,\theta) := \mathrm{W}_2^2(P_\theta\#\gamma_Y, P_\theta\#\gamma_{Z^{(j)}}).$$

By sum, all of the previous results will hold, with the only technical point being path differentiability, which is stable by sum ([10], Corollary 4). Note that this extension is also valid for a Monte-Carlo approximation of $\mathcal{E}$, replacing $\mathcal{E}$ with $\mathcal{E}_p$ in the barycentre formulation.

4.5. **A Result for Decreasing Learning Rates.** In [18], Davis et al. show the convergence of *decreasing-step* noised SGD of a function F under certain conditions. Our goal is to apply their Theorem 4.2 to $F \in \{\mathcal{E},\mathcal{E}_p\}$ with the following SGD scheme:

$$\begin{aligned} Y^{(t+1)} &= Y^{(t)} - \alpha^{(t)}\varphi(Y^{(t)},\theta^{(t+1)}) + \alpha^{(t)}a\varepsilon^{(t+1)},\\ &(Y^{(0)},(\theta^{(t)})_{t\in\mathbb{N}},(\varepsilon^{(t)})_{t\in\mathbb{N}}) \sim \nu\otimes\mu^{\otimes\mathbb{N}}\otimes\eta^{\otimes\mathbb{N}}, \end{aligned} \tag{4.10}$$

where as before, $\mu\in\{\varpi,\varpi_p\}$ for $\mathcal{E},\mathcal{E}_p$ respectively, ν is the distribution of the initial position $Y^{(0)}$, $\eta := \mathcal{N}(0,I_D)$ is the noise distribution (it could be chosen more generally, but we attempt to stay close to our previous formalism for simplicity), and finally the learning rate sequence $(\alpha^{(t)})$ verify:

$$\forall t\in\mathbb{R},\ \alpha^{(t)}\geq 0, \quad \sum_{t=0}^{+\infty}\alpha^{(t)} = +\infty, \text{ and } \sum_{t=0}^{+\infty}(\alpha^{(t)})^2 < +\infty.$$

Theorem 4.3. *Consider $(Y^{(t)})$ a trajectory of (4.10) for $\mu \in \{\mathfrak{w}, \mathfrak{w}_p\}$ respectively, assume that it is almost-surely bounded. Then the sequence $F(Y^{(t)})$ is almost-surely convergent for $F \in \{\mathcal{E}, \mathcal{E}_p\}$ respectively, and almost-surely, any subsequential limit of $(Y^{(t)})$ belongs to the set of Clarke critical points of F.*

Proof. We verify assumptions C.1, C.2, C.3 and D.1, D.2 of [18], allowing us to apply Theorem 4.2. To begin with, the sequence $(\alpha^{(t)})$ was chosen to verify assumption C.1, and Theorem 4.3 explicitly assumes C.2. Regarding C.3, the simple choice of independent noise verifies the martingale difference assumption trivially.

For D.1, we need to show that the set of non-critical points of $\mathcal{E}$ and $\mathcal{E}_p$ are dense in $\mathbb{R}^D$. For $\mathcal{E}$, we can use Corollary 3.1, which implies that critical points Y of $\mathcal{E}$ necessarily verify $\sum y_k = \sum_k z_k$, since $\sum_k S_{k,l}^{Y,Z} = I_d$, or are points of non-differentiability, which implies belonging to $\mathcal{U}$ (see (2.9)) In particular, critical points are necessarily within a union of two strict subspaces of $\mathbb{R}^D$, whose complementary is dense in $\mathbb{R}^D$. For $\mathcal{E}_p$, the property is easily verified using its decomposition into (non-trivial) quadratics on cells Proposition 2.2.

For D.2, we leverage [18] Lemma 5.2 along with the fact that $\mathcal{E}$ and $\mathcal{E}_p$ are path-differentiable (Section 4.3), which shows that D.2 is verified. □

The assumption that the trajectories are almost-surely bounded can be seen as a discrete version of [32] Assumption 4.4: A1), which Li and Moosmüller require to prove the convergence of their decreasing-step SGD scheme for SW between absolutely continuous measures. While this assumption is theoretical costly, numerically we observe that the measure support Y remains bounded. Nevertheless, lifting this assumption would be of substantial mathematical interest.

5. Numerical Experiments

This section illustrates the optimisation properties of $\mathcal{E}$ and $\mathcal{E}_p$ with several numerical experiments. Section 5.1 studies the optimisation of $\mathcal{E}_p$ using the BCD algorithm described in Algorithm 1, which offers insights on the cell structure of $\mathcal{E}_p$ (Section 2.3). Section 5.2.1 focuses on stochastic gradient descent Algorithm 2 and showcases various SGD trajectories on $\mathcal{E}$ and $\mathcal{E}_p$ for different learning rates, noise levels or numbers of projections, as well as the Wasserstein error along iterations. All the convergence curves shown throughout our experiments also showcase margins of error, computed by repeating the experiments several times, and corresponding to the 30% and 70% quantiles of the experiment.

In order to assess the quality of a position $Y^{(t)}$, perhaps the most germane metric is the Wasserstein distance: $\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$, which is why we will study the 2-Wasserstein error of BCD and SGD trajectories in this section. Unfortunately, this metric is not quite comparable for different dimensions d, notably because $\|(1, \cdots, 1)\|_2^2 = d$. We shall attempt to compensate this phenomenon by using $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ instead, which makes the metric more comparable for measures on spaces of different dimensions.

5.1. Empirical study of Block Coordinate Descent on $\mathcal{E}_p$.

In this section, we shall focus on studying the optimisation properties of the $\mathcal{E}_p$ landscape using the BCD algorithm (Algorithm 1). This method leverages the cell structure of $\mathcal{E}_p$ (see Section 2.3), by moving from cell to cell by computing the minimum of their associated quadratics (see the discussion in Section 3.2.4). By Theorem 3.2,

all local optima of $\mathcal{E}_p$ are stable cell optima, i.e. fixed points of the BCD, which summarises briefly the ties between BCD and the optimisation properties of $\mathcal{E}_p$. As for the numerical implementation, Algorithm 1 was implemented in Python with Numpy [25] using the closed-form formulae for the updates.

5.1.1. *Illustration in 2D.*
Dataset and implementation details. We start by setting a simple 2D measure γ_Z with a support of only two points represented with stars in Figure 5. The measure weights are taken as uniform. We fix sequences of p projections $(\theta_1, \cdots, \theta_p)$ for $p \in \{3, 10, 30, 100\}$ respectively. We then draw 100 BCD schemes with different initial positions $Y^{(0)} \in \mathbb{R}^{2\times 2}$, drawn with independent standard Gaussian entries. We take a stopping criterion threshold of 10^{-5} (see Algorithm 1), and limit to 500 iterations.

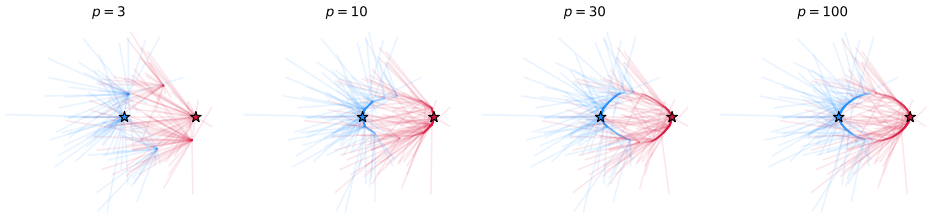
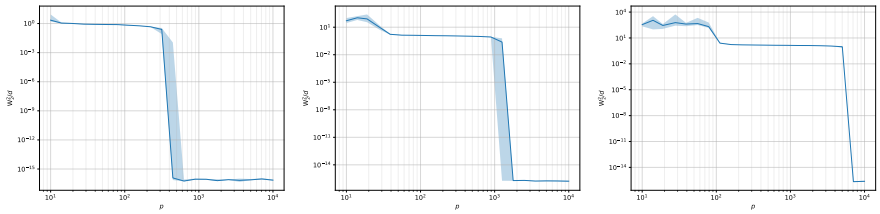


FIGURE 5. BCD on $\mathcal{E}_p$ with different initial positions $Y^{(0)}$, with fixed projections (first sample). Each of the two points of the trajectory $Y^{(t)} = (y_1^{(t)}, y_2^{(t)})$ is coloured with respect to the point of the original measure γ_Z to which they converge.

In the case $p = 3$, we observe on Figure 5 four points which correspond to strict local optima, and the schemes appear to have a comparable probability of converging towards each of them. Note that these points are essentially the same as the ones represented in Figure 2 for $p = 3$, but that they depend on the projection sample. Between the two projection realisations, we observe that these local optima change locations. The cases $p \in \{10, 30, 100\}$ also exhibit strict local optima, however they appear to be decreasingly likely to be converged towards. For $p = 30$ and $p = 100$, notice that most trajectories end up on the same ellipsoid arcs towards the solution Z, and further remark that these arcs strongly resembles the trajectories of SGD schemes on $\mathcal{E}$ for small learning rates (see Figure 9 in Section 5.2).

5.1.2. *Wasserstein convergence of BCD schemes on $\mathcal{E}_p$.*
Final Wasserstein error of BCD Schemes. For a dimension $d \in \{10, 30, 100\}$ and $n = 20$ points, the original measure γ_Z, $Z \in \mathbb{R}^{n\times d}$ is sampled once for all with independent standard Gaussian entries. Then, for varying numbers of projections p, we draw a starting position $Y^{(0)} \in \mathbb{R}^{n\times d}$ with entries that are uniform on $[0, 1]$; and draw p projections as input to the BCD algorithm. We set the stopping criterion threshold as $\varepsilon = 10^{-5}$ and the maximum iterations to 1000. In order to produce Figure 6, we record the normalised 2-Wasserstein discrepancy $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z)$ at the final iteration T for 10 realisations for each value of p and d.

As a first estimation of the difficulty of optimising $\mathcal{E}_p$, we consider the evolution - as p increases - of final W_2^2 errors of BCD schemes. The results of the experiments



(A) $d = 10$ (B) $d = 30$ (C) $d = 100$

FIGURE 6. We consider BCD schemes with different amounts of projections p, and with an original measure γ_Z comprised of $n = 10$ points in dimension $d \in \{10, 30, 100\}$, which is fixed as a standard Gaussian realisation for each value of d. The stopping threshold was chosen as $\varepsilon = 10^{-5}$, and we plot the final Wasserstein errors $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^T}, \gamma_Z)$ at the final iteration T. For each set of values for the parameters, we perform 10 realisations with different initialisations $Y^{(0)}$ (drawn with uniform $[0, 1]$ entries), and different projections $(\theta_1, \cdots, \theta_p)$.

presented in Figure 6 suggests the existence of a phase transition between an insufficient and a sufficient amount of projections. For instance, in the case $d = 10$, there appears to be a cutoff around $p = 400$, under which all the BCD realisations converge towards strict local optima, and past which we observe convergence up to numerical precision.

Probability of convergence of BCD schemes. We can investigate further this empirical cutoff phenomenon by estimating the probability of convergence of a BCD algorithm. This probability is loosely related to the difficulty of optimising the landscape $\mathcal{E}_p$, since a high probability of BCD convergence indicates either a small number of strict local optima, or that their corresponding cells are extremely small and seldom reached in practice. For varying numbers of projections p and dimensions d, we run 100 realisations of BCD schemes. Each sample draws a target measure γ_Z, $Z \in \mathbb{R}^{n\times d}$ with independent standard Gaussian entries and $n = 10$ points, as well as its initialisation $Y^{(0)} \in \mathbb{R}^{n\times d}$ with entries that are uniform on $[0, 1]$ and p projections. Every BCD scheme has a stopping threshold of $\varepsilon = 10^{-5}$ and a maximum of 1000 iterations. We consider that a sample scheme has converged (towards the global optimum γ_Z) if $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z) < 10^{-5}$, which allows us to compute an empirical probability of convergence for each value of (p, d).

The findings in Figure 7 indicate that the W_2^2 error cutoffs from Figure 6 have a probabilistic counterpart: the probability of converging to a global optimum transitions from almost 0 to almost 1 relatively suddenly (in the logarithmic scale). We can conjecture that this drop in optimisation difficulty is tied to the number of iterations needed for the convergence of SGD schemes on $\mathcal{E}$, especially given the similar behaviour for the W_2^2 error in Figure 13.

5.2. **Empirical study of SGD on $\mathcal{E}$ and $\mathcal{E}_p$.**

General numerical implementation. In order to perform gradient descent on $\mathcal{E}$ or $\mathcal{E}_p$, we compute the gradient (4.3) using Pytorch's [39] Stochastic Gradient Descent optimiser, which back-propagates gradients through the loss $w_\theta := Y \mapsto$

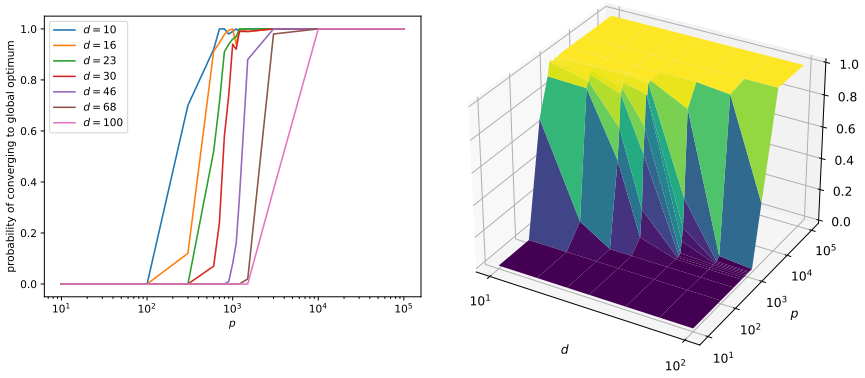


FIGURE 7. Given a number of projections p, we run 100 BCD trials with different initial positions (with entries drawn as uniform on $[0,1]$), projections and target measure supported by $Z \in \mathbb{R}^{n\times d}$, with $n = 10$ points in different dimensions $d \in [10, 100]$, where Z is drawn with independent standard Gaussian entries. At the final iteration T, we determine whether the optimum is global by a threshold criterion: $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z) < 10^{-5}$ and compute an empirical probability of convergence.

$\mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z)$, which we compute using the 1D Wasserstein solver from Python Optimal Transport [22].

5.2.1. *Illustration in 2D.*
2D dataset and implementation details. We define a 2D spiral dataset with the measure γ_Z, $Z = (z_1, \cdots, z_{10})^T \in \mathbb{R}^{10\times 2}$ with $z_k = \frac{2k}{10}\left(\cos\left(2k\pi/10\right), \sin\left(2k\pi/10\right)\right)^T$, and $k \in [\![1, 10]\!]$. The initial position $Y^{(0)}$ is fixed and remains the same across realisations. For schemes on $\mathcal{E}$, the projections $\theta^{(t)} \sim \mathfrak{o}$ are fixed beforehand and are the same across experiments. For every realisation of a scheme on $\mathcal{E}_p$, p unique projections $(\theta_1, \cdots, \theta_p)$ are drawn, then the projections $(\theta^{(t)})$ for the iterations are drawn from these p fixed projections. For noised schemes, the only variable that is drawn at every sample is the noise $(\varepsilon^{(t)})$. Note that the associated energy landscapes are extremely similar to those illustrated in Section 2.6 and in particular in Figure 2.

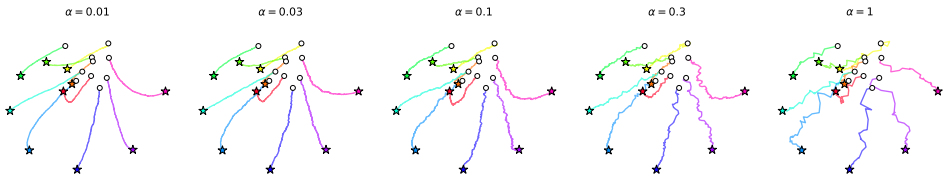


FIGURE 8. SGD trajectories on $\mathcal{E}$ for different learning rates α. All the trajectories are computed using the same projection sequence $(\theta^{(t)})$.

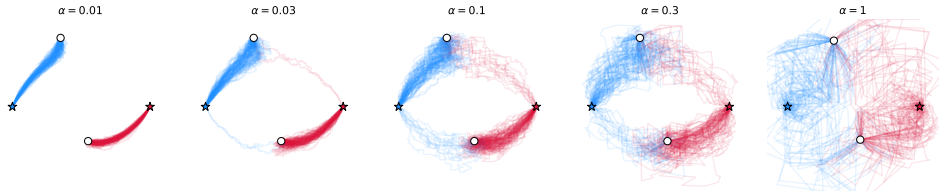


FIGURE 9. SGD trajectories on $\mathcal{E}$ for different learning rates α. For each value of α, 100 samples are drawn with different projections $(\theta^{(t)})$, and for each realisation, each of the two points of the trajectory is coloured with respect to the point of the original measure γ_Z (represented by stars) to which they converge. The initial position $Y^{(0)}$ is represented by circles.

Figure 8 and Figure 9 illustrate the convergence of SGD schemes on $\mathcal{E}$ towards the original measure γ_Z, for different learning rate α (provided that α is under a divergence threshold). Theorem 4.1 allowed us only to expect a convergence to a *solution of a Clarke Differential Inclusion* on $\mathcal{E}$ (4.5), yet in practice we seem to have convergence to a global optimum. Furthermore, Theorem 4.1 shows that the interpolated SGD trajectories are approximately solutions of the DI $\dot{X}(t) \in -\partial_C \mathcal{E}(X(t))$, which, assuming that the trajectory stays in $\mathcal{U}$, amounts to $\dot{X}(t) + \nabla\mathcal{E}(X(t)) = 0$, which is exactly the Euclidean Gradient Flow of $\mathcal{E}$, as discussed in more detail in Remark 4.1. This illustration suggests that the SGD schemes approach the gradient flow (4.7) as $\alpha \longrightarrow 0$, whereas Theorem 4.1 predicts a (weak) convergence towards the set of solutions of the DI (4.5), which is equal to the gradient flow provided that the initial position $Y^{(0)}$ belongs to the differentiability set of $\mathcal{E}$ (see Remark 4.1 for details). Note that higher learning rates lead to a "noisier" trajectory, which may impede upon the quality of the assignment. This shows that there is a trade-off: lower values of α allow for a better approximation of the (or a) gradient flow of $\mathcal{E}$ and potentially a more precise final position Y and assignment τ, however a larger value of α yields a substantially faster convergence.

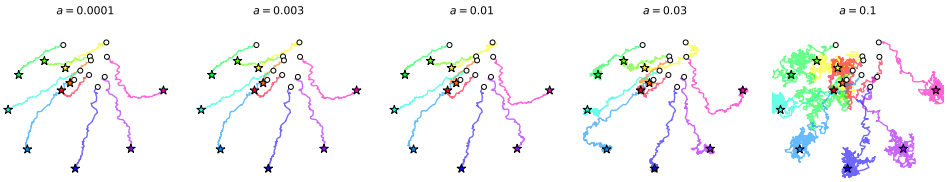


FIGURE 10. SGD trajectories on $\mathcal{E}$ for different noise levels a. All the trajectories are computed using the same projection sequence $(\theta^{(t)})$. The learning rate is fixed at $\alpha = 0.3$.

Figure 10 presents a case where noised SGD schemes on $\mathcal{E}$ "converge" whatever the noise level to a global optimum of $\mathcal{E}$. Note that the additive noise causes the scheme to oscillate around a solution, with a movement akin to Brownian motion with a scale tied to αa. Theorem 4.2 shows that such schemes converge (as the step approaches 0) to *Clarke critical points* of $\mathcal{E}$, which could theoretically be a

saddle point of strict local optimum. In this experiment, we observe convergence to a global optimum.

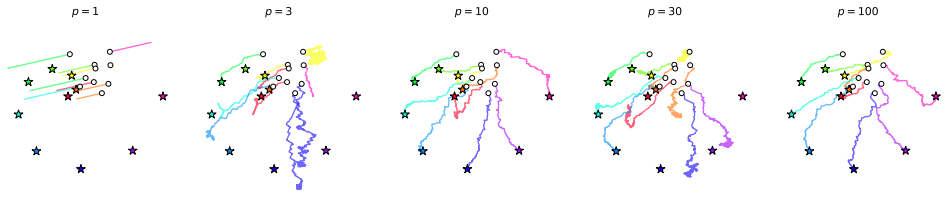


FIGURE 11. SGD schemes on $\mathcal{E}_p$ for different number of projections p. The learning rate is fixed at $\alpha = 0.3$.

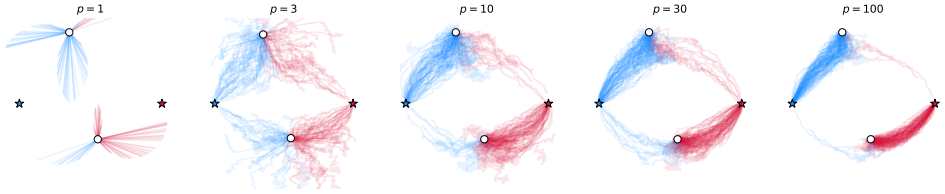


FIGURE 12. SGD schemes on $\mathcal{E}_p$ for different number of projections p. For each value of p, 100 samples are drawn with different projections $(\theta_1, \cdots, \theta_p)$. For each realisation, each of the two points of the trajectory is coloured with respect to the point of the original measure γ_Z (represented by stars) to which they converge. The initial position $Y^{(0)}$ is represented by circles. The learning rate is fixed at $\alpha = 0.03$.

Figure 11 illustrates that SGD schemes on $\mathcal{E}_p$ may converge to strict local optima, which is to be expected, given how numerous they may be (see the discussion in Section 2.6 and Figure 2 therein). For $p = 1$, entire lines are local optima, and for $p = 3$ and $p = 30$, we also observe convergence to strict local optima. Notice that for a large value of p such as $p = 100 \gg d = 2$, we have similar trajectories in Figure 12 compared to the $\mathcal{E}$ counterpart in Figure 9 ($\alpha = 0.03$). This observation suggests a stronger property than our results on the approximation of $\mathcal{E}$ by $\mathcal{E}_p$: uniform convergence in Theorem 2.3 and a weak link between critical points Theorem 3.3. To be precise, this illustration could allow one to hope for a result on the high probability for the proximity of SGD schemes on $\mathcal{E}_p$ and on $\mathcal{E}$ as $p \longrightarrow +\infty$, perhaps with conditions on the sequence of projections $(\theta^{(t)})$.

5.2.2. *Wasserstein convergence of SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$.*
SGD on $\mathcal{E}$. The original measure γ_Z, $Z \in \mathbb{R}^{n \times d}$ is sampled once for all with independent standard Gaussian entries. For each value of the parameter of interest (the learning rate α or the dimension d respectively), 10 realisations of the SGD schemes are computed with a different initial position $Y^{(0)}$, drawn with independent entries uniform on $[0, 1]$, and different projections $(\theta^{(t)})$. The SGD stopping criterion threshold (see Algorithm 2) is set as negative, in order to always end at the maximum number of iterations, 10^6. For the experiment with varying learning

rates α, we consider measures with $n = 20$ points in dimension $d = 10$. For the experiment with varying dimensions d, we still take $n = 20$ and use the learning rate $\alpha = 10$.

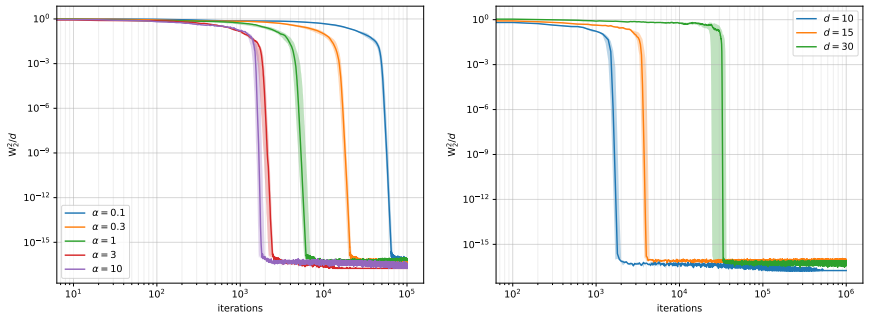


FIGURE 13. Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for SGD iterations $Y^{(t)}$ on $\mathcal{E}$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. Left: different learning rates α for points in dimension $d = 10$. Right: different dimensions with $\alpha = 10$ (right).

In Figure 13, we observe that the SGD schemes converge towards the true measure γ_Z up to numerical precision, which corresponds to a stronger convergence than the one predicted by Theorem 4.1. The number of iterations needed for convergence obviously depends on the learning rate α, which notably can be chosen larger than $n/2$, which is a case that does not fall under the conditions for Theorem 4.1. However, in this particular experiment, the SGD schemes diverged as soon as $\alpha \geq 30$, which could suggest that limiting oneself to $\alpha \preceq n$ is reasonable. The dimension d increases significantly the number of iterations required for convergence, furthermore we observe a transition from high W_2^2 error to low error, which is relatively sudden in logarithmic space. These first studies invites an in-depth analysis of the amount of iterations needed to reach convergence, which we propose in Figure 16. The final $\frac{1}{d}\mathrm{W}_2^2$ error does not seem to depend significantly on the dimension d, which provides empirical grounds for the $1/d$ normalisation choice.

Noised SGD on $\mathcal{E}$. Figure 14 shows the Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for the noised SGD iterations on $\mathcal{E}$. The numerical setup is the same as above, with the addition of the noise $a\alpha\varepsilon^{(t)}$ at each iteration, where $\varepsilon^{(t)}$ has independent standard Gaussian entries, a is the noise level and α is the learning rate (set to $\alpha = 10$). This noise is drawn differently for each SGD scheme. For the experiment with different dimensions, the noise level is taken as $a = 10^{-4}$.

The noised SGD scheme errors oscillate around a certain level which depends on the noise level, as the trajectories from Figure 10 suggest: we observed Brownian-like motion around the target points. Note that the error begins falling drastically past the same iteration threshold, albeit with a higher variance across samples for higher noise levels. At a fixed noise level, the final $\frac{1}{d}\mathrm{W}_2^2$ still depends on the noise level, despite the $1/d$ normalisation. Empirically, the final W_2^2 error seems to be smaller than the noise level a, which is reassuring since the noise is entry-wise of law $\mathcal{N}(0, a^2\alpha^2)$, where α is the learning rate.

SGD on $\mathcal{E}_p$. Figure 15 also illustrates the Wasserstein error along iterations but this time for $\mathcal{E}_p$. The general SGD setup and initial measure γ_Z remain unchanged

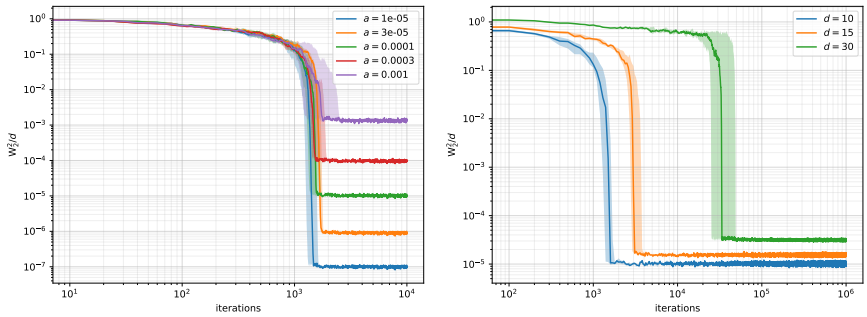


FIGURE 14. Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for noised SGD iterations $Y^{(t)}$ on $\mathcal{E}$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. The noise is additive standard Gaussian, scaled by the learning rate $\alpha = 10$ times the noise level a. Left: different noise levels a for points in dimension $d = 10$. Right: different dimensions with $a = 10^{-4}$.

compared to the schemes on $\mathcal{E}$ (with also a learning rate of $\alpha = 10$ in particular). In order to handle the projections $(\theta^{(t)})$, for each sample we draw p independent projections $(\theta_1, \cdots, \theta_p)$, then select the $(\theta^{(t)})$ by drawing uniformly amongst these p projections. Given this sequence of projections $(\theta^{(t)})$, the SGD algorithm is then exactly the same as for $\mathcal{E}$.

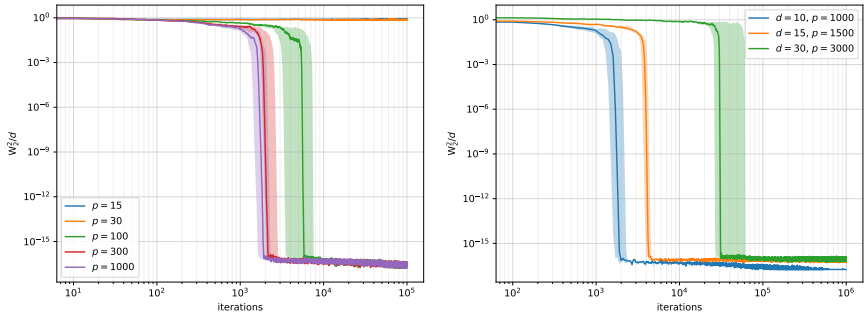


FIGURE 15. Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for SGD iterations $Y^{(t)}$ on $\mathcal{E}_p$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. The p projections in $\mathcal{E}_p$ are drawn randomly for each sample. Left: different noise levels a for points in dimension $d = 10$. Right: different dimensions with $\alpha = 10$.

For SGD schemes on $\mathcal{E}_p$ with small values of projections p, we do not have convergence to $\gamma_{Y^{(t)}} = \gamma_Z$. Intuitively, this could be understood as the approximation $\mathcal{E}_p \approx \mathcal{E}$ being too rough, allowing for an excessive amount of numerically attainable strict local optima. This is illustrated in Figure 2 in a simple case: with $p = 3$ in dimension 2, the landscape presents numerous strict local optima that lie within large basins. However, it is notable that for p large enough ($p \geq 10d = 100$), we *do* observe convergence to $\gamma_{Y^{(t)}} = \gamma_Z$ up to numerical precision. This convergence

happens in fewer iterations as p increases, and with a smaller variance with respect to the projection samples. This suggests a stronger mode of convergence of $\mathcal{E}_p$ towards $\mathcal{E}$, as hinted at before in Figure 2 and Figure 12.

Quantifying the impact of the dimension. For different values of the number of points n and the dimension d, we run 10 samples of SGD on $\mathcal{E}$ for an original measure γ_Z drawn with standard Gaussian entries (re-drawn for each sample this time). The SGD schemes are done without additive noise, and with a learning rate of $\alpha = 10$. In order to save computation time, the SGD stopping threshold is taken as $\beta = 10^{-5}$ (see Algorithm 2). For each sample, the initial position $Y^{(0)}$ is drawn with entries that are uniform on $[0, 1]$. Our goal is to estimate the number of iterations required for the convergence of the SGD schemes: to this end, we define convergence as the first step t such that $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z) < 10^{-5}$.

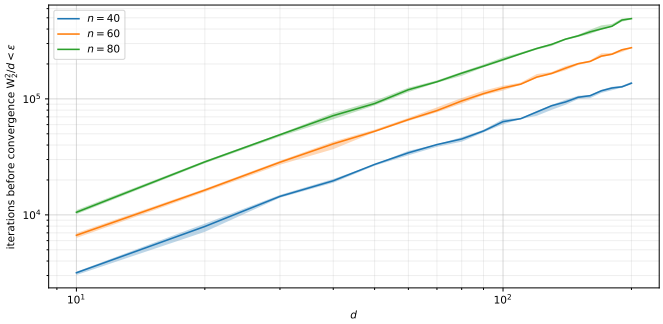


FIGURE 16. Amount of iterations required for convergence of a SGD scheme $Y^{(t)}$ of learning rate $\alpha = 10$ on $\mathcal{E}$. Here convergence is defined as the first step t such that $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z) < 10^{-5}$. For each set of parameters (number of points n and dimension d), 10 trials are done with γ_Z, $Z \in \mathbb{R}^{n\times d}$ drawn at random (uniform $[0, 1]^{n\times d}$).

Figure 16 (cautiously) suggests that the number of iterations required for convergence its proportional to $d^{1.25}$ (where convergence means that $\frac{1}{d}\mathrm{W}_2^2$ falls below ε). Note that the exponent on d does not seem to depend on n. Obviously, the factor in front of $d^{1.25}$ depends on the number of points n, the learning rate α and the convergence threshold ε. This superlinear rule remains fairly prohibitive for large Machine Learning models, which can typically have d and n both in excess of 10^6.

6. Conclusion and Outlook

Throughout this paper, we have investigated the properties of the Sliced Wasserstein (SW) distance between discrete measures, namely the function $\mathcal{E} : Y \longmapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, where Y and Z are supports with n points in dimension d. Due to the intractability of the expectation in $\mathcal{E}$, we introduced its Monte-Carlo empirical counterpart $\mathcal{E}_p$, computed as an average over p directions. In Section 2, we showed and reminded regularity results on $\mathcal{E}$ and $\mathcal{E}_p$: they are locally-Lipschitz and differentiable on certain open sets of full measure. Leveraging the fact that $\mathcal{E}_p$ is piece-wise quadratic, we showed additional regularity results, and finally showed

that the convergence of $\mathcal{E}_p$ to $\mathcal{E}$ (as $p \longrightarrow +\infty$) is almost-surely uniform on any fixed compact. Section 3 furthers the study of the optimisation landscapes at hand by presenting properties of the critical points of $\mathcal{E}$ and $\mathcal{E}_p$ (points of differentiability will null gradient), and a convergence of such points of $\mathcal{E}_p$ to those of $\mathcal{E}$ as $p \longrightarrow +\infty$ (in a certain sense). In Section 4, we put these theoretical results in a more practical context by showing that one can apply the SGD convergence results of [6] to our optimisation landscapes. Finally, we illustrate and study these convergence results in Section 5 through numerical experiments.

Further work would be welcome on the cells of $\mathcal{E}_p$ (see Section 2.3), in particular the law of their size given a fixed configuration $\mathbf{m}$ and their probability of being stable are still open problems, and would have strong consequences in practical applications such as the convergence of BCD (Algorithm 1). The main difficulty stems from the link between statistical properties of the cells to the so-called Gaussian Orthant Probabilities, which can be broadly defined as the probability of a non-standard Gaussian Vector to be in the positive quadrant $\mathbb{R}^d_+$. This probability is unfortunately not tractable in high dimensions, and its estimation is a field of research in itself [3].

Another core limitation of our work concerns the practicality of our results on SGD convergence (Section 4). Firstly, typical applications use more advanced optimisation methods, such as SGD with momentum or ADAM, which our theory does not encompass yet. Secondly, as mentioned in the introduction, practical applications actually minimise *through* $\mathcal{E}$, which is to say a loss function $F : u \longmapsto \mathrm{SW}_2^2(T_u\#\mu, \nu)$ with respect to the parameters u of a model $x \longmapsto T_u(x)$ of the input data $x \sim \mu$. Minimising F through SGD (stochastically on the projections $\theta \sim \varpi$, the input data $x \sim \mu$ and the true data $y \sim \nu$) is beyond the scope of this paper, and we leave this generalisation for future work.

Acknowledgements. We thank Anna Korba and Quentin Mérigot for their helpful discussions and comments on the optimisation results. We would also like to thank Antoine Chambaz for his valuable assistance with empirical processes. Finally, we are grateful for the numerous suggestions of anonymous reviewers that substantially improved this paper.

This research was funded, in part, by the Agence nationale de la recherche (ANR), through the SOCOT project (ANR-23-CE40-0017), and the PEPR PDE-AI project (ANR-23-PEIA-0004).

References

[1] Hana Alghamdi, Mairead Grogan, and Rozenn Dahyot, *Patch-based colour transfer with optimal transport*, 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 2019, pp. 1–5.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, *Wasserstein generative adversarial networks*, Proceedings of the 34th International Conference on Machine Learning (Doina Precup and Yee Whye Teh, eds.), Proceedings of Machine Learning Research, vol. 70, PMLR, 06–11 Aug 2017, pp. 214–223.

[3] Dario Azzimonti and David Ginsbourger, *Estimating orthant probabilities of high-dimensional Gaussian vectors with an application to set estimation*, J. Comput. Graph. Statist. **27** (2018), no. 2, 255–267. MR 3816262

[4] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar, *Stochastic gradient descent for barycenters in wasserstein space*, (2022).

[5] Erhan Bayraktar and Gaoyue Guo, *Strong equivalence between metrics of Wasserstein type*, Electronic Communications in Probability **26** (2021).

[6] 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** (2022), no. 3, 1117–1147.
[7] ———, *Stochastic subgradient descent escapes active strict saddles on weakly convex functions*, Mathematics of Operations Research (2023).
[8] Xin Bing, Florentina Bunea, and Jonathan Niles-Weed, *The sketched Wasserstein distance for mixture distributions*, arXiv preprint arXiv:2206.12768 (2022).
[9] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization **18** (2007), no. 2, 556–572.
[10] Jérôme Bolte and Edouard Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning*, Mathematical Programming **188** (2021), 19–51.
[11] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister, *Sliced and Radon Wasserstein barycenters of measures*, Journal of Mathematical Imaging and Vision **51** (2015), no. 1, 22–45.
[12] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister, *Blind video temporal consistency*, ACM Transactions on Graphics (TOG) **34** (2015), no. 6, 1–9.
[13] Nicolas Bonnotte, *Unidimensional and evolution methods for optimal transportation.*, PhD Thesis, Paris 11 (2013).
[14] Frank H Clarke, *Optimization and nonsmooth analysis*, SIAM, 1990.
[15] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems **26** (2013).
[16] Damek Davis and Dmitriy Drusvyatskiy, *Proximal methods avoid active strict saddles of weakly convex functions*, Foundations of Computational Mathematics **22** (2022), no. 2, 561–606.
[17] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang, *Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization*, 2023.
[18] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee, *Stochastic subgradient method converges on tame functions*, Foundations of computational mathematics **20** (2020), no. 1, 119–154.
[19] Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing, *Generative modeling using the sliced Wasserstein distance*, 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3483–3491.
[20] Richard Mansfield Dudley, *The speed of mean Glivenko-Cantelli convergence*, The Annals of Mathematical Statistics **40** (1969), no. 1, 40–50.
[21] 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).
[22] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer, *POT: Python optimal transport*, Journal of Machine Learning Research **22** (2021), no. 78, 1–8.
[23] Aude Genevay, Gabriel Peyré, and Marco Cuturi, *Learning generative models with Sinkhorn divergences*, International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1608–1617.
[24] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, *Improved training of Wasserstein GANs*, Advances in neural information processing systems **30** (2017).
[25] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al., *Array programming with numpy*, Nature **585** (2020), no. 7825, 357–362.
[26] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour, *A sliced Wasserstein loss for neural texture synthesis*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9412–9420.

[27] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan, *How to escape saddle points efficiently*, International conference on machine learning, PMLR, 2017, pp. 1724–1732.

[28] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan, *On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points*, Journal of the ACM (JACM) **68** (2021), no. 2, 1–29.

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, *Progressive growing of GANs for improved quality, stability, and variation*, arXiv preprint arXiv:1710.10196 (2017).

[30] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde, *Sliced Wasserstein auto-encoders*, International Conference on Learning Representations, 2019.

[31] S. Levy, F. Hirsch, and G. Lacombe, *Elements of functional analysis*, Graduate Texts in Mathematics, Springer New York, 2012.

[32] Shiying Li and Caroline Moosmueller, *Measure transfer via stochastic slicing and matching*, (2023).

[33] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter, *Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions*, Proceedings of the 36th International Conference on Machine Learning (Kamalika Chaudhuri and Ruslan Salakhutdinov, eds.), Proceedings of Machine Learning Research, vol. 97, PMLR, 09–15 Jun 2019, pp. 4104–4113.

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

[35] Quentin Mérigot, Filippo Santambrogio, and Clément Sarrazin, *Non-asymptotic convergence bounds for Wasserstein approximation using point clouds*, Advances in Neural Information Processing Systems **34** (2021), 12810–12821.

[36] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli, *Approximate bayesian computation with the sliced-Wasserstein distance*, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 5470–5474.

[37] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli, *Statistical and topological properties of sliced probability divergences*, Advances in Neural Information Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds.), vol. 33, Curran Associates, Inc., 2020, pp. 20802–20812.

[38] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau, *Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance*, Advances in Neural Information Processing Systems (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, *Pytorch: An imperative style, high-performance deep learning library*, Advances in Neural Information Processing Systems (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[40] G. Peyré and M. Cuturi, *Computational optimal transport*, Foundations and Trends in Machine Learning **51** (2019), no. 1, 1–44.

[41] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot, *Wasserstein barycenter and its application to texture mixing*, 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, Springer, 2012, pp. 435–446.

[42] Filippo Santambrogio, *Optimal transport for applied mathematicians*, Birkäuser, NY **55** (2015), no. 58-63, 94.

[43] Eloi Tanguy, Rémi Flamary, and Julie Delon, *Reconstructing discrete measures from projections. consequences on the empirical sliced Wasserstein distance*, arXiv preprint arXiv:2304.12029 (2023).

[44] Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau, *Wasserstein loss for image synthesis and restoration*, SIAM Journal on Imaging Sciences **9** (2016), no. 4, 1726–1755.

[45] Joel A Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of computational mathematics **12** (2012), no. 4, 389–434.

[46] Aad W Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.
[47] Jean-Philippe Vial, *Strong and weak convexity of sets and functions*, Mathematics of Operations Research **8** (1983), no. 2, 231–259.
[48] Cédric Villani, *Optimal transport : old and new / cédric villani*, Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 2009 (eng).
[49] Seiichiro Wakabayashi, *Remarks on semi-algebraic functions*, January 2008, Online Notes.
[50] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. Paudel, and L. Van Gool, *Sliced Wasserstein generative models*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Los Alamitos, CA, USA), IEEE Computer Society, jun 2019, pp. 3708–3717.
[51] Jiaqi Xi and Jonathan Niles-Weed, *Distributional convergence of the sliced Wasserstein process*, Advances in Neural Information Processing Systems **35** (2022), 13961–13973.
[52] 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).

Appendix A.

A.1. **Proof of the Central Limit Theorem for Discrete SW.** Let $c_{\mathcal{K}} := \sup_{Y\in K} \|Y\|_{\infty,2}$. Consider the class of functions $\mathcal{F} := \{f_Y |\ Y \in \mathcal{K}\}$, where we define $f_Y := \theta \longmapsto \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z)$ to fit the notation style of empirical processes. By Proposition 2.1, we have for any $\theta \in \mathbb{S}^{d-1}$ and $Y, Y' \in \mathcal{K}$:

$$|f_Y(\theta) - f_{Y'}(\theta)| \leq 2n(2c_{\mathcal{K}} + c_{\mathcal{K}} + \|Z\|_{\infty,2})\|Y - Y'\|_{\infty,2},$$

where we chose the neighbourhood $B_{\|\cdot\|_{\infty,2}}(0, 2c_{\mathcal{K}}) \supset \mathcal{K}$. In particular, the Lipschitz constant $\kappa := 2n(3c_{\mathcal{K}} + \|Z\|_{\infty,2})$ is $\mathfrak{o}$-integrable (in this case, it is constant in θ), which allows us to apply Example 19.7 from [46], which implies that the family $\mathcal{F}$ is $\mathfrak{o}$-Donsker.

To recall the definition of the property that $\mathcal{F}$ is $\mathfrak{o}$-Donsker, we recall some standard concepts and notations from empirical processes ([46] Section 19.2), within our specific case of application. For f a function from $\mathbb{S}^{d-1}$ to $\mathbb{R}$ that is $\mathfrak{o}$-square-integrable, we introduce the real random variable

$$\mathfrak{o}_p f := \frac{1}{p}\sum_{i=1}^{p} f(\theta_i),\ (\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p},$$

which is meant as a Monte-Carlo approximation of the true expectation $\mathfrak{o}f := \int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\mathfrak{o}(\theta)$. We shall study the distribution of the scaled approximation error, defined as the real random variable $\mathbb{G}_p f := \sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f)$. The central limit theorem shows that $\mathbb{G}_p f$ converges in law to a centred Gaussian (in our case the functions $f \in \mathcal{F}$ are trivially $\mathfrak{o} - L^2$), and our goal is instead to show the uniform convergence of the *random process*

$$\mathbb{G}_p := \{\sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f),\ f \in \mathcal{F}\}.$$

To study the convergence of the sequence of processes $(\mathbb{G}_p)_{p\in\mathbb{N}^*}$, we introduce the space $\ell^\infty(\mathcal{F})$ of the bounded functions $z : \mathcal{F} \longrightarrow \mathbb{R}$, equipped with the norm $\|z\|_\infty = \sup_{f\in\mathcal{F}}(|z(f)|)$. The class of functions $\mathcal{F}$ is said to be $\mathfrak{o}$-Donsker if each process $\mathbb{G}_p$ is in $\ell^\infty(\mathcal{F})$ (which is to say that it has bounded trajectories), and if the sequence $(\mathbb{G}_p)_{p\in\mathbb{N}^*}$ converges in law (in the sense of the topology of $\ell^\infty(\mathcal{F})$) towards a tight[3] process $\mathbb{G}$. Due to the usual multivariate Central Limit Theorem, this process $\mathbb{G}$ is necessarily the $\mathfrak{o}$-*Brownian Bridge*, which is to say the centred Gaussian process indexed on $\mathcal{F}$ such that $\mathrm{Cov}[\mathbb{G}f, \mathbb{G}f'] = \mathfrak{o}(ff') - (\mathfrak{o}f)(\mathfrak{o}f')$.

We have now shown that our error process $\mathbb{G}_p$ converges in law (in $\ell^\infty(\mathcal{F})$) towards the Gaussian process $\mathbb{G}$, i.e.

$$\{\sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f),\ f \in \mathcal{F}\} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\ \ell^\infty(\mathcal{F})} \mathbb{G}. \tag{A.1}$$

Seeing our energy $\mathcal{E}_p : Y \longmapsto \mathfrak{o}_p f_Y$ as a process on $\mathcal{K}$, we can identify the indexes $\{f_Y,\ Y \in \mathcal{K}\}$ with $\mathcal{K}$ itself, yielding a convergence of processes on $\mathcal{K}$. To be precise, we define the space $\ell^\infty(\mathcal{K})$ as the space of bounded functions $z : \mathcal{K} \longrightarrow \mathbb{R}$, equipped with the infinite norm. Consider the map

$$\phi := \left\{ \begin{array}{ccc} \ell^\infty(\mathcal{F}) & \longrightarrow & \ell^\infty(\mathcal{K}) \\ z & \longmapsto & \left\{ \begin{array}{ccc} \mathcal{K} & \longrightarrow & \mathbb{R} \\ Y & \longmapsto & z(f_Y) \end{array} \right. \end{array} \right. ,$$

[3]Since this technical notion is not essential for our result, we refer to [46] page 260 for a complete presentation.

since $\mathcal{F} := \{f_Y |\ Y \in \mathcal{K}\}$, it is well defined, and it is continuous, since it is linear and verifies for any $z \in \ell^\infty(\mathcal{F})$, $\|\phi(z)\|_{\ell^\infty(K)} = \sup_{Y\in\mathcal{K}} |z(f_Y)| = \sup_{f\in\mathcal{F}} |z(f)| = \|z\|_{\ell^\infty(\mathcal{F})}$. By the continuous mapping theorem (see [46] Theorem 18.11 for its use on stochastic processes), we can apply ϕ to the convergence in law in (A.1), yielding

$$\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathcal{K})} \phi(\mathbb{G}), \tag{A.2}$$

where the process $\phi(\mathbb{G})$ is the centred Gaussian process on $\mathcal{K}$ with the covariance structure $\mathrm{Cov}\phi(\mathbb{G})[Y, Y'] = \varpi(f_Y f_{Y'}) - (\varpi(f_Y))(\varpi(f_{Y'}))$. Note that $\mathcal{E}(Y) = \varpi f_Y$ with our notations.

With the continuous mapping theorem (again [46] Theorem 18.11), we can apply the continuous map $\|\cdot\|_{\ell^\infty(\mathcal{K})} : \ell^\infty(\mathcal{K}) \longrightarrow \mathbb{R}$, yielding a uniform convergence result

$$\sqrt{p}\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} \|\phi(\mathbb{G})\|_{\ell^\infty(\mathcal{K})}. \tag{A.3}$$

A.2. **Computing $\mathcal{E}$, W_2^2 and $\mathcal{E}_p$ in a simple case.**
Computing $\mathcal{E}$. We work in polar coordinates, writing

$$\theta = \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}, \text{ and } y = \begin{pmatrix} u \\ v \end{pmatrix} = r \begin{pmatrix} \cos\psi \\ \sin\psi \end{pmatrix}.$$

By symmetry of the problem, we can assume $\psi \in [0, \pi/2]$ (i.e. the top-right quadrant $u \geq 0, v \geq 0$). Now let $\psi \in [0, 2\pi[$, let us compute $\mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z)$. Since we project in 1D, computing this slice amounts to sorting $(\theta^T y_1, \theta^T y_2)$ and $(\theta^T z_1, \theta^T z_2)$. Let $\tau_Z^\theta \in \mathfrak{S}_2$ such that $\theta^T y_{\tau_Y^\theta(1)} \leq \theta^T y_{\tau_Y^\theta(2)}$ and similarly $\theta^T z_{\tau_Z^\theta(1)} \leq \theta^T z_{\tau_Z^\theta(2)}$. We always have

$$\mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z) = \frac{1}{2}\left(\left(\theta^T(y_{\tau_Y^\theta(1)} - z_{\tau_Z^\theta(1)})\right)^2 + \left(\theta^T(y_{\tau_Y^\theta(2)} - z_{\tau_Z^\theta(2)})\right)^2\right).$$

We split the integral depending on the values of τ_Y^θ and τ_Z^θ, which vary depending on the angle of the projection ϕ. We begin with τ_Y^θ:

$$\theta^T y_1 \geq \theta^T y_2 \Longleftrightarrow \cos\phi\cos\psi + \sin\phi\sin\psi \geq 0 \Longleftrightarrow \phi \in [\psi - \pi/2, \psi + \psi/2] + 2\pi\mathbb{Z}. \tag{A.4}$$

The equation for τ_Z^θ is much simpler:

$$\theta^T z_1 \geq \theta^T z_2 \Longleftrightarrow -\sin\phi \geq 0 \Longleftrightarrow \phi \in [\pi, 2\pi] + 2\pi\mathbb{Z}. \tag{A.5}$$

We divide a period of 2π in four quadrants corresponding to the four possibilities for $(\tau_Y^\theta, \tau_Z^\theta)$. Since we assume $\psi \in [0, \pi/2]$, we can write this simply as:

$$\begin{aligned}
\mathcal{E}(Y) = &\frac{1}{4\pi}\int_{-\pi}^{\psi-\pi/2} \left(\left(\theta^T(y_1 - z_2)\right)^2 + \left(\theta^T(y_2 - z_1)\right)^2\right) \mathrm{d}\phi \\
&+ \frac{1}{4\pi}\int_{\psi-\pi/2}^{0} \left(\left(\theta^T(y_2 - z_2)\right)^2 + \left(\theta^T(y_1 - z_1)\right)^2\right) \mathrm{d}\phi \\
&+ \frac{1}{4\pi}\int_{0}^{\psi+\pi/2} \left(\left(\theta^T(y_2 - z_1)\right)^2 + \left(\theta^T(y_1 - z_2)\right)^2\right) \mathrm{d}\phi \\
&+ \frac{1}{4\pi}\int_{\psi+\pi/2}^{\pi} \left(\left(\theta^T(y_1 - z_1)\right)^2 + \left(\theta^T(y_2 - z_2)\right)^2\right) \mathrm{d}\phi.
\end{aligned}$$

Elementary trigonometric integration yields

$$\mathcal{E}(Y) = \frac{r^2}{2} + \frac{1}{2} - \frac{2}{\pi}\left(r\cos\psi + r\psi\sin\psi\right) = \frac{u^2+v^2}{2} + \frac{1}{2} - \frac{2}{\pi}\left(u + v\,\mathrm{Arctan}\,\frac{v}{u}\right), \tag{A.6}$$

which holds for $\psi \in [0, \pi/2]$. By symmetry, we obtain the following expression for any $(u, v) \in \mathbb{R}^2$ (recall that we stack the vectors in Y line by line):

$$\mathcal{E}\begin{pmatrix} u & v \\ -u & -v \end{pmatrix} = \frac{u^2+v^2}{2} + \frac{1}{2} - \frac{2}{\pi}\left(|u| + |v|\,\mathrm{Arctan}\left|\frac{v}{u}\right|\right). \tag{A.7}$$

In the general case, dimension d would require the use of d-dimensional spherical coordinates, making the equations (A.4) and (A.5) intractable. Furthermore, generalising to n points would separate the integral into $(n!)^2$ parts, losing all hopes of tractability and legibility.

Computing W_2^2. In the case $n = 2$, the Kantorovich LP formulation of the Wasserstein distance can be written as:

$$\min_{a\in[0,1]} \sum_{k,l\in[\![1,2]\!]} \pi_{k,l}(a)\|y_k - z_l\|_2^2, \quad \text{with } \pi(a) := \frac{1}{2}\begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix}.$$

Substituting $y_1 = \begin{pmatrix} u \\ v \end{pmatrix}$, $y_2 = \begin{pmatrix} -u \\ -v \end{pmatrix}$, $z_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$, $z_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ yields:

$$\mathrm{W}_2^2(\gamma_Y, \gamma_Z) = \min_{a\in[0,1]} \left(u^2 + (v+1)^2 - 4av\right) = u^2 + (|v|-1)^2.$$

Computing $\mathcal{E}_p$. For simplicity, in the following we will only consider $\theta \in \mathbb{S}^{d-1}$ such that the $\theta^T y_k$ are distinct, and such that the $\theta^T z_k$ are also distinct. We will express the cases for the values of the sortings τ_Y^θ and τ_Z^θ in a different (yet equivalent) manner.

We have $\tau_Y^\theta = I$ if $\theta^T y_1 < \theta^T y_2$ and $\tau_Y^\theta = (2,1)$ otherwise. Then $\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I$ if $\tau_Y^\theta = \tau_Z^\theta$, and $\tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1} = (2,1)$ otherwise. The system

$$\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I \iff \begin{cases} \theta^T y_1 < \theta^T y_2 \text{ and } \theta^T z_1 < \theta^T z_2 \\ \text{or} \\ \theta^T y_2 < \theta^T y_1 \text{ and } \theta^T z_2 < \theta^T z_1 \end{cases}$$

can be simplified, yielding:

$$\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I \iff \left(\theta\theta^T(z_2 - z_1)\right)^T (y_2 - y_1) > 0. \tag{A.8}$$

(A.8) is a linear equation in Y. Additionally, (A.8) only depends on $y_2 - y_1 = -2y$, which makes our symmetrical simplification inconsequential. Plugging in the specific point values yields a more explicit definition of the cells. We write the condition on $y \in \mathbb{R}^2$, since $Y = (y, -y)^T$.

$$\mathcal{C}_{\mathbf{m}} = \left\{ y \in \mathbb{R}^2 \mid \forall i \in [\![1,p]\!],\ -\mathrm{sign}\left[\theta_i^T \begin{pmatrix} 0 \\ 1 \end{pmatrix}\ \theta_i^T y\right] = +1 \text{ if } \mathbf{m}_i = I, \text{ else} - 1 \right\}. \tag{A.9}$$

Equation (A.9) describes $\mathcal{C}_{\mathbf{m}}$ as an intersection of p half-planes of $\mathbb{R}^2$, thus it is a polytope. Note that we use strict inequalities, which lifts configuration ambiguities, and implies that the $(\mathcal{C}_{\mathbf{m}})_{\mathbf{m}\in\mathfrak{S}_2^p}$ are disjoint, and that the union of their closure is $\mathbb{R}^2$.

Straightforward computation yields

$$\operatorname*{argmin}_{X\in\mathbb{R}^{n\times d}} q_{\mathbf{m}}(X) = (A^{-1}(B_{1,1}^{\mathbf{m}} z_1 + B_{1,2}^{\mathbf{m}} z_2), A^{-1}(B_{2,1}^{\mathbf{m}} z_1 + B_{2,2}^{\mathbf{m}} z_2)),$$

where $A := \frac{1}{p}\sum_{i=1}^{p} \theta_i\theta_i^T$ and $B^{\mathbf{m}}_{k,l} := \frac{1}{p}\sum_{\substack{i=1\\ \mathbf{m}_i(k)=l}}^{p} \theta_i\theta_i^T$.

Note that our $n = 2$ setting, we have the simplifications $B^{\mathbf{m}}_{1,2} = A - B^{\mathbf{m}}_{1,1}$, $B^{\mathbf{m}}_{2,1} = B^{\mathbf{m}}_{1,2}$ and $B^{\mathbf{m}}_{1,1} = B^{\mathbf{m}}_{2,2}$. Furthermore, $B^{\mathbf{m}}_{k,l}$ is (up to a factor), a Monte-Carlo estimation of $S^{Y,Z}_{k,l}$ (see Corollary 3.1).

A.3. **Discrete Wasserstein stability.** Consider the following generic discrete Kantorovich problem, given weights $\alpha \in \Sigma_n$ and $\beta \in \Sigma_m$ and a generic cost matrix $C \in \mathbb{R}_+^{n\times m}$:

$$\mathrm{W}(\alpha, \beta; C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C, \tag{A.10}$$

where $\Pi(\alpha, \beta, C)$ is the set of $n \times m$ matrices π with non-negative entries such that $\pi\mathbb{1} = \alpha$ and $\pi^T\mathbb{1} = \beta$.

Lemma A.1 (Stability of the Wasserstein cost)**.** *Let $\alpha, \overline{\alpha}, \beta, \overline{\beta} \in \Sigma_n$, and $C, \overline{C} \in \mathbb{R}_+^{n\times n}$, such that the weights verify $\alpha, \overline{\alpha}, \beta, \overline{\beta} > 0$ entry-wise. Then:*

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C - \overline{C}\|_\infty + \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1). \tag{A.11}$$

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C - \overline{C}\|_F + \|C\|_F \left(\|\alpha - \overline{\alpha}\|_2 + \|\beta - \overline{\beta}\|_2\right), \tag{A.12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Note that this result is a generalisation of [8], Theorem 2 (they assumes that the cost matrices are pairwise distances, which amount to the W_1 case), but requires the weights to have positive entries (as opposed to non-negative entries).

Proof. We split the difference in two terms:

$$\begin{aligned}\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| &\leq \left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\alpha, \beta; \overline{C})\right| =: \mathrm{I}\\ &\quad + \left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; C)\right| =: \mathrm{II}\end{aligned}$$

— *Step 1*: Controlling I Using the primal formulation

We use Equation (A.10): let $\overline{\pi}^*$ optimal for $\mathrm{W}(\alpha, \beta, \overline{C})$. In particular, $\overline{\pi}^*$ is admissible for the problem $\mathrm{W}(\alpha, \beta, C)$. We have

$$\begin{aligned}\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\alpha, \beta; \overline{C}) &= \min_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C - \min_{\pi\in\Pi(\alpha,\beta)} \pi \cdot \overline{C}\\ &\leq \overline{\pi}^* \cdot C - \overline{\pi}^* \cdot \overline{C} = \sum_{i=1}^{n}\sum_{j=1}^{m} \overline{\pi}^*_{i,j}(C_{i,j} - \overline{C}_{i,j})\\ &\leq \|C - \overline{C}\|_\infty \sum_{i=1}^{n}\sum_{j=1}^{m} \overline{\pi}^*_{i,j} = \|C - \overline{C}\|_\infty,\end{aligned}$$

where the property $\pi \in \Pi(\alpha, \beta)$ implied $\sum_{i,j} \overline{\pi}^*_{i,j} = 1$. By using the same argument symmetrically, we obtain

$$\mathrm{I} \leq \|C - \overline{C}\|_\infty.$$

— *Step 2*: Controlling the dual variables

Consider the Legendre dual problem associated to (A.10):

$$\mathrm{W}(\alpha, \beta; C) = \sup_{\substack{f\in\mathbb{R}^n, g\in\mathbb{R}^m\\ f\oplus g\leq C}} f^T\alpha + g^T\beta. \tag{A.13}$$

Let f^*, g^* optimal for the dual formulation, our objective is to bound this dual solution in a set which depends on C. First, notice that the value and constraints remain unchanged if we replace (f^*, g^*) with $(f^* - t\mathbb{1}, g^* + t\mathbb{1})$ for $t \in \mathbb{R}$, which allows us to assume $f_1^* = 0$. We now leverage the complementary slackness property (which characterises the primal-dual optimality conditions for this linear problem, see [40] Section 3.3): for any π^* optimal for the primal problem (A.10), we have the implication

$$\pi^*_{i,j} \neq 0 \Longrightarrow f_i^* + g_j^* = C_{i,j}.$$

The primal constraints imply that $\sum_j \pi^*_{1,j} = \alpha_1 > 0$ and that $\pi^* \geq 0$ entry-wise, there exists $j_1 \in [\![1, m]\!]$ such that $\pi^*_{1,j_1} \neq 0$. Using the complementary slackness implication, we obtain $0 + g^*_{j_1} = C_{1,j_1}$. We now use the dual constraint $f^* \oplus g^* \leq C$ at $i = 1$ to show that $\forall j \in [\![1, m]\!],\ g_j^* \leq C_{1,j}$. This allows us to find a lower-bound on f^*: since $\forall i \in [\![2, n]\!],\ \sum_j \pi^*_{i,j} = \alpha_i > 0$, thus there exists a $j_i \in [\![1, m]\!]$ such that $\pi^*_{i,j_i} \neq 0$, yielding $f_i^* + g^*_{j_i} = C_{i,j_i}$, then since $g^*_{j_i} \leq C_{1,j_i}$, this yields the lower-bound $f_i^* \geq C_{i,j_i} - C_{1,j_i}$. For an upper-bound on f^*, we use the dual constraint at (i, j_1): we have $f_i^* + g^*_{j_1} \leq C_{i,j_1}$, then we use $g^*_{j_1} = C_{1,j_1}$ proved earlier to show $f_i^* \leq C_{i,j_1} - C_{1,j_1}$. At this point, we have the following control on f_i^*:

$$f_1^* = 0, \quad \forall i \in [\![2, n]\!],\ C_{i,j_i} - C_{1,j_i} \leq f_i^* \leq C_{i,j_1} - C_{1,j_1}.$$

Regarding g^*, we already have $\forall j \in [\![1, m]\!],\ g_j \leq C_{1,j}$. For a lower bound, since $\sum_i \pi_{i,j} = \beta_j > 0$, there exists $i_j \in [\![1, n]\!]$ such that $\pi^*_{i_j,j} \neq 0$, so by complementary slackness $f^*_{i_j} + g^*_j = C_{i_j,j}$, thus by the upper-bound on f^* we have if $i_j \neq 1$ that $g^*_j \geq C_{i_j,j} - C_{i_j,j_1} + C_{1,j_1}$. If $i_j = 1$ then $f^*_{i_j} = 0$ and $g^*_j = C_{i_1,j}$. Our control on g^* is the following:

$$\forall j \in [\![1, m]\!], \begin{cases} C_{i_j,j} - C_{i_j,j_1} + C_{1,j_1} \leq g_j^* \leq C_{1,j} & \text{if } i_j \neq 1 \\ g_j^* = C_{i_1,j} & \text{if } i_j = 1 \end{cases}.$$

We summarise our bounds in the following (weaker) statement, which holds thanks to the condition $C \geq 0$ (entry-wise):

$$\|f^*\|_\infty \leq \|C\|_\infty, \quad \|g^*\|_\infty \leq \|C\|_\infty.$$

— *Step 3*: Bounding II using the dual formulation

Let f^*, g^* optimal for the dual formulation (A.13) of $\mathrm{W}(\alpha, \beta, C)$, which by Step 2 we can choose to verify $\|f^*\|_\infty \leq \|C\|_\infty$ and $\|g^*\|_\infty \leq \|C\|_\infty$. In particular, (f^*, g^*) is admissible for the dual formulation of $\mathrm{W}(\overline{\alpha}, \overline{\beta}, C)$.

$$\begin{aligned}
\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; C) &= \max_{f \oplus g \leq C} f^T\alpha + g^T\beta - \max_{\overline{f} \oplus \overline{g} \leq C} \overline{f}^T\overline{\alpha} + \overline{g}^T\overline{\beta} \\
&\leq (f^*)^T\alpha + (g^*)^T\beta - (f^*)^T\overline{\alpha} - (g^*)^T\overline{\beta} \\
&= (f^*)^T(\alpha - \overline{\alpha}) + (g^*)^T(\beta - \overline{\beta}) \\
&\leq \|f^*\|_\infty\|\alpha - \overline{\alpha}\|_1 + \|g^*\|_\infty\|\beta - \overline{\beta}\|_1 \\
&\leq \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1).
\end{aligned}$$

By symmetry, we obtain $|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; C)| \leq \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1)$.

— *Step 4*: Wrapping up

By Step 1 and Step 3 combined we conclude:

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \mathrm{I} + \mathrm{II} \leq \|C - \overline{C}\|_\infty + \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1).$$

— Details for the proof of (A.12)

For the first term, to get the Frobenius norm $\|C - \overline{C}\|_F$ instead of the infinite norm, it suffices to use that $\|M\|_\infty \leq \|M\|_F$.

For the second term, note that the penultimate inequality of Step 3 can also be written with the Cauchy-Schwarz inequality, yielding $\|C\|_F(\|\alpha - \overline{\alpha}\|_2 + \|\beta - \overline{\beta}\|_2)$, where the upper-bound on $\|f^*\|_2$ and $\|g^*\|_2$ by $\|C\|_F$ are obtained using the element-wise bounds on f^* and g^* from Step 2. □

A.4. **Proof of Theorem 3.3 and convergence rate.** The proof of Theorem 3.3 requires matrix concentration technicalities. In the following, $\|\cdot\|_{\text{op}}$ denotes the $\|\cdot\|_2$-induced operator norm on $\mathbb{R}^{d\times d}$, and $S_d(\mathbb{R})$ denotes the space of symmetric $d\times d$ matrices. We write $\preceq$ for the Loewner order of positive semi-definite symmetric matrices ($A \preceq B$ means that $B-A$ is positive semi-definite). We recall the following Hoeffding inequality.

Theorem A.1 (Matrix Hoeffding Inequality, [45], Theorem 1.3)**.**

Let $q \in \mathbb{N}^*$, $(X_i)_{i\in[\![1,q]\!]}$ independent random variables with values in $S_d(\mathbb{R})$, such that $\mathbb{E}[X_i] = 0$. Suppose that $\forall i \in [\![1,q]\!]$, $\exists A_i \in S_d(\mathbb{R}) : X_i^2 \preceq A_i^2$. Let $\sigma^2 := \left\|\sum_i A_i^2\right\|_{\text{op}}$, then for any $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^q X_i\right\|_{\text{op}} \geq t\right) \leq d\exp\left(-\frac{t^2}{8\sigma^2}\right).$$

We deduce from Theorem A.1 the following lemma, where the X_i follow a uniform law on $\Theta \subset \mathbb{S}^{d-1}$.

Lemma A.2 (Hoeffding applied to $\theta \sim \mathcal{U}(\Theta)$)**.**

Let $(\theta_i)_{i\in[\![1,q]\!]}$, independent random vectors following the uniform law on $\Theta \subset \mathbb{S}^{d-1}$, where Θ is σ-measurable with $\sigma(\Theta) > 0$. Let $S_\Theta := \frac{1}{s_\Theta}\int_\Theta \theta\theta^T \mathrm{d}\sigma(\theta)$, where $s_\Theta := \sigma(\Theta)$. S_Θ is the covariance matrix of $\theta \sim \mathcal{U}(\Theta)$. Let $\eta \in]0,1[$ and $t > 0$. Then with probability exceeding $1-\eta$ we have

$$q \geq \frac{32\log(d/\eta)}{t^2} \Longrightarrow \left\|\frac{1}{q}\sum_{i=1}^q \theta_i\theta_i^T - S_\Theta\right\|_{\text{op}} \leq t.$$

In the case $\Theta = \mathbb{S}^{d-1}$, the condition $q \geq \dfrac{8\log(d/\eta)}{t^2}$ is sufficient.

Proof. The idea is to apply Theorem A.1 to $X_i := \frac{1}{q}\theta_i\theta_i^T - \frac{1}{q}S_\Theta$. First, by definition, $\mathbb{E}[X_i] = 0$.

We now find $A \in S_d^+(\mathbb{R})$ such that $X_i^2 \preceq A$. Let $u \in \mathbb{S}^{d-1}$, we compute:

$$u^T X_i^2 u = \frac{1}{q^2}\left(u^T\theta_i\theta_i^T u - u^T\theta_i\theta_i^T S_\Theta u - u^T S_\Theta\theta_i\theta_i^T u + u^T S_\Theta^2 u\right) \leq \left(\frac{1+\|S_\Theta\|_{\text{op}}}{q}\right)^2.$$

Moreover, $\|S_\Theta\|_{\text{op}} \leq 1$, since

$$\forall u \in \mathbb{S}^{d-1},\ u^T S_\Theta u = \frac{1}{s_\Theta}\int_\Theta u^T\theta\theta^T u \mathrm{d}\sigma(\theta) \leq \frac{1}{s_\Theta}\int_\Theta 1\mathrm{d}\sigma(\theta) = 1.$$

In conclusion $X_i^2 \preceq \frac{4}{q^2}I$. Using the notations of Theorem A.1, we compute $\sigma^2 = 4/q$, and apply the Matrix Hoeffding inequality with $\Delta := \sum_i X_i = \frac{1}{q}\sum_i \theta_i\theta_i^T - S_\Theta$. It follows that for any $t > 0$, $\mathbb{P}(\|\Delta\|_{\text{op}} \geq t) \leq d\exp\left(-\frac{qt^2}{32}\right)$. In order to have the

event $\|\Delta\|_{\text{op}} \leq t$ with probability exceeding $1-\eta$, it is therefore sufficient that $\eta \geq d\exp\left(-\frac{qt^2}{32}\right)$, which is equivalent to $q \geq \frac{32\log(d/\eta)}{t^2}$.

In the case $\Theta = \mathbb{S}^{d-1}$, one has $S_\Theta = I/d$, and a finer Loewner upper-bound can be established, since

$$u^T X_i^2 u = \frac{1}{q^2}\left(u^T\theta_i\theta_i^T u - \frac{2}{d}u^T\theta_i\theta_i^T u + \frac{1}{d^2}\right) \leq \left(\frac{1-\frac{1}{d}}{q}\right)^2 \leq \frac{1}{q^2},$$

and thus $\sigma^2 = 1/q$. This yields the Hoeffding inequality $\mathbb{P}\left(\|\Delta\|_{\text{op}} \geq t\right) \leq d\exp\left(-\frac{qt^2}{8}\right)$, which in turn provides the announced weaker condition on q. □

With this tool at hand, we now prove a quantitative concentration result:

Theorem A.2 (Concentration of cell optima)**.**

Let $\mathbf{m} = (\sigma_1, \cdots, \sigma_p)$ be a fixed matching configuration (see Section 2.3) and let $(\theta_i)_{i\in[\![1,p]\!]} \sim \mathfrak{v}^{\otimes p}$ (uniform on $\mathbb{S}^{d-1}$). We introduce the following notations and variables:

- *For $(k,l) \in [\![1,n]\!]^2$, let $q_{k,l} := \#\{i \in [\![1,p]\!] \mid k = \sigma_i(l)\}$;*
- *Let $\overline{c}_Z := \max_{l\in[\![1,n]\!]} \|z_l\|_2$;*
- *Let $\varepsilon \in]0, \frac{4}{3}n\overline{c}_Z]$;*
- *Let $\eta \in]0,1[$.*

Assume the following:

- $(H_q):\ \forall (k,l) \in [\![1,n]\!]^2,\ q_{k,l} \geq \overline{q}$ *or* $q_{k,l} < \underline{q}$, *with* $1 \leq \underline{q} \leq \overline{q} \leq p$;
- $(H_1):\ p \geq \dfrac{697d^2n^2\overline{c}_Z^2\log(3d/\eta)}{\varepsilon^2}$;
- $(H_2):\ \overline{q} \geq \dfrac{512d^2\overline{c}_Z^2\log(3dnn^+/\eta)}{\varepsilon^2}$; $\quad n^+ := \max_{k\in[\![1,n]\!]} \#\{l \in [\![1,n]\!] \mid q_{k,l} \geq \overline{q}\}$;
- $(H_3):\ \underline{q} \leq \dfrac{\varepsilon}{8dn^-\overline{c}_Z}\, p$; $\quad n^- := \max_{k\in[\![1,n]\!]} \#\{l \in [\![1,n]\!] \mid q_{k,l} \leq \underline{q}\}$;
- $(H_4):\ p \geq \dfrac{8d^2n^2\overline{c}_Z^2\log(6n^2/\eta)}{\varepsilon^2}$.

Then with probability exceeding $1-\eta$, writing $Y^* := \operatorname{argmin}_{Y'\in\mathbb{R}^{n\times d}} q_{\mathbf{m}}(Y')$, we have

$$\forall k \in [\![1,n]\!],\ \left\| y_k^* - \sum_{l=1}^n S_{k,l} z_l \right\|_2 \leq \varepsilon, \tag{A.14}$$

where the normalized conditional covariance matrices $S_{k,l}$ are defined in Corollary 3.1 (we omit the Y^*, Z exponent here for legibility).

Proof. — Step 1: Re-writing (3.4).

Remind that the matching configuration $\mathbf{m}$ is fixed here. Let $Y^* := \operatorname{argmin}_{Y'\in\mathbb{R}^{n\times d}} q_{\mathbf{m}}(Y')$ and $k \in [\![1,n]\!]$. By (3.4), we have

$$y_k^* = A^{-1}\left(\frac{1}{p}\sum_{i=1}^p \theta_i\theta_i^T z_{\sigma_i(k)}\right),\ \text{with } A = \frac{1}{p}\sum_{i=1}^p \theta_i\theta_i^T.$$

Let $I_{k,l} := \{i \in [\![1,p]\!] \mid \sigma_i(k) = l\}$. Since the σ_i are permutations, we have $[\![1,p]\!] = \bigcup_{l=1}^{n} I_{k,l} = \bigcup_{k=1}^{n} I_{k,l}$ and $k \neq k' \Rightarrow I_{k,l} \cap I_{k',l} = \varnothing$; $l \neq l' \Rightarrow I_{k,l} \cap I_{k,l'} = \varnothing$. We re-order the sum:

$$\frac{1}{p}\sum_{i=1}^{p} \theta_i \theta_i^T z_{\sigma_i(k)} = \sum_{l=1}^{n} \frac{1}{p} \sum_{i \in I_{k,l}} \theta_i \theta_i^T z_l = \sum_{l=1}^{n} \frac{q_{k,l}}{p} B_{k,l} z_l,$$

where $q_{k,l} := \#I_{k,l}$ and $B_{k,l} := \dfrac{1}{q_{k,l}} \displaystyle\sum_{i \in I_{k,l}} \theta_i \theta_i^T$. This invites the definition of the matrix $R = (r_{k,l})$, $r_{k,l} := \frac{q_{k,l}}{p}$, which is bi-stochastic by construction.

— *Step 2*: Separating the terms in y_k^*.

We will see later that the empirical covariance matrix A concentrates towards the covariance matrix of $\theta \sim \varpi$, which is I/d. In order to quantify the impact of this concentration on y_k^*, we introduce the error term: $\delta A^- := A^{-1} - dI$.

A similar concentration will be observed for $B_{k,l}$, but the θ_i in the sum are *selected* such that $i \in I_{k,l}$. Recall that since we project in 1D, the permutations σ_i arise from a sorting problem, namely $\sigma_i = \tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1}$, where we recall that τ_Y^θ is a permutation sorting the numbers $(y_1^T\theta, \cdots, y_n^T\theta)$.

By definition, we have $\sigma_i(k) = l \iff \theta_i \in \Theta_{k,l} = \left\{\theta \in \mathbb{S}^{d-1} \mid \tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k) = l\right\}$, where we omit again the Y, Z exponent on $\Theta_{k,l}$ for legibility.

Since the θ_i in $B_{k,l}$ are drawn under the condition $\theta_i \in \Theta_{k,l}$, we study the concentration $B_{k,l} \approx C_{k,l}$, where $C_{k,l} := \frac{1}{d\varpi(\Theta_{k,l})} S_{k,l}$. In order to quantify this approximation, we define the error term $\delta B_{k,l} := B_{k,l} - C_{k,l}$. Similarly, the $r_{k,l} := \frac{q_{k,l}}{p}$ are Monte-Carlo approximations of $\varpi(\Theta_{k,l})$, which leads to the definition $\delta r_{k,l} := r_{k,l} - \varpi(\Theta_{k,l})$.

We may now separate the terms in the result from Step 1:

$$\begin{aligned} y_k^* &= (dI + \delta A^-) \left(\sum_{l=1}^{n} r_{k,l} (\underbrace{C_{k,l} + \delta B_{k,l}}_{B_{k,l}}) z_l \right) \\ &= \underbrace{d \sum_{l=1}^{n} \varpi(\Theta_{k,l}) C_{k,l} z_l}_{v} + \underbrace{\delta A^- \left(\sum_{l=1}^{n} r_{k,l} B_{k,l} z_l \right)}_{\delta v_1} + \underbrace{d \sum_{\substack{l=1 \\ q_{k,l} \geq \overline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l}_{\delta v_2} \\ &\quad + \underbrace{d \sum_{\substack{l=1 \\ q_{k,l} < \underline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l}_{\delta v_3} + \underbrace{d \sum_{l=1}^{n} \delta r_{k,l} C_{k,l} z_l}_{\delta v_4}. \end{aligned}$$

The separation of the terms in the second equality arises from (H_q), formulated in the theorem. Observe that the first term v is exactly $\Psi(Y^*)$, with Ψ defined in Section 3.1.2. Our objective is to provide conditions under which $\forall i \in \{1,2,3,4\}$, $\|\delta v_i\|_2 \leq \varepsilon/4$ with probability exceeding $1-\eta$. To that end, we let $\varepsilon > 0$ and $\eta \in]0,1[$.

— *Step 3*: Condition for $\|\delta v_2\|_2 \leq \dfrac{\varepsilon}{4}$.

First of all, note that if the sum defining δv_2 is empty, the condition holds trivially almost-surely. In the following, we suppose that the sum has at least one non-zero term. We have from Step 2,

$$\|\delta v_2\|_2 = \left\| d \sum_{\substack{l=1 \\ q_{k,l} \geq \overline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l \right\|_2 \leq d\overline{c}_Z \sum_{\substack{l=1 \\ q_{k,l} \geq \overline{q}}}^{n} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}}.$$

Let the shorthands $n_k^+ := \# J_k^+$ and $J_k^+ := \{l \in [\![1,n]\!] \mid q_{k,l} \geq \overline{q}\}$. We upper-bound the right term by $\sum_{l \in J_k^+} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}} \leq \sum_{l \in J_k^+} r_{k,l} \max_{l \in J_k^+} \|\delta B_{k,l}\|_{\mathrm{op}} \leq \max_{l \in J_k^+} \|\delta B_{k,l}\|_{\mathrm{op}}$.

For $l \in J_k^+$, by Lemma A.2, we have $\|\delta B_{k,l}\|_{\mathrm{op}} \leq t$ with probability exceeding $1 - \eta/(3nn_k^+)$ provided that $q_{k,l} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}$. Since the probability of $\bigcup_{l \in J_k^+} \{\|\delta B_{k,l}\|_{\mathrm{op}} > t\}$ can be upper bounded by the sum of the probabilities of each of the n_k^+ terms, it is upper bounded by $\eta/(3n)$. Therefore, writing the event $\{\forall l \in J_k^+,\ \|\delta B_{k,l}\|_{\mathrm{op}} \leq t\}$ as the complementary of this union, we conclude that it holds with probability exceeding $1 - \eta/(3n)$, provided that

$$\forall l \in J_k^+,\ q_{k,l} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}.$$

A sufficient condition for this last assumption to hold is $(H_2^k):\ \overline{q} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}$. Applying this result to $t := \frac{\varepsilon}{4d\overline{c}_Z}$, and by letting $n^+ := \max_{k \in [\![1,n]\!]} n_k^+$, a sufficient condition to have $\|\delta v_2\|_2 \leq \frac{\varepsilon}{4}$ with probability exceeding $1 - \eta/(3n)$ is

$$(H_2): \quad \overline{q} \geq \frac{512 d^2 \overline{c}_Z^2 \log(3dnn^+/\eta)}{\varepsilon^2}.$$

— *Step 4*: Condition for $\|\delta v_3\|_2 \leq \frac{\varepsilon}{4}$.

With a computation analogous to Step 3, we write

$$\|\delta v_3\|_2 = \left\| d \sum_{\substack{l=1 \\ q_{k,l} < \underline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l \right\|_2 \leq d\overline{c}_Z \sum_{l \in J_k^-} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}},$$

where, like in Step 3, we define $n_k^- := \# J_k^-$ and $J_k^- := \{l \in [\![1,n]\!] \mid q_{k,l} \leq \underline{q}\}$. If $n_k^- = 0$ then the objective holds almost-surely, thus we suppose $n_k^- \geq 1$. In this setting, the $q_{k,l}$ are small, thus we have little control over $\|\delta B_{k,l}\|_{\mathrm{op}}$, which can be upper bounded by 2.

Leveraging the condition $q_{k,l} \leq \underline{q}$, which holds for $l \in J_k^-$, we have $r_{k,l} = q_{k,l}/p \leq \underline{q}/p$. In order to have $\|\delta v_3\|_2 \leq \frac{\varepsilon}{4}$ almost-surely, it is sufficient to have $(H_3^k):\ \underline{q} \leq \frac{\varepsilon}{8dn_k^- \overline{c}_Z}\, p$. Again, with $n^- := \max_{k \in [\![1,n]\!]} n_k^-$, we obtain the sufficient condition:

$$(H_3): \quad \underline{q} \leq \frac{\varepsilon}{8dn^- \overline{c}_Z}\, p.$$

— *Step 5*: Condition for $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$.

By definition, $\delta v_4 = d\sum_{l=1}^{n} \delta r_{k,l} C_{k,l} z_l$, then $\|\delta v_4\|_2 \leq \bar{c}_Z d \sum_{l=1}^{n} |\delta r_{k,l}| \|C_{k,l}\|_{\mathrm{op}}$. We use the upper-bound $\|C_{k,l}\|_{\mathrm{op}} \leq 1$ (observe that $\|C_{k,l}\|_{\mathrm{op}}$ can be made as close to 1 as desired by choosing $\Theta_{k,l}$ as a very small portion of the sphere). In order to have $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$, it is sufficient to have $\forall l \in [\![1,n]\!],\ |\delta r_{k,l}| \leq \frac{\varepsilon}{4d n \bar{c}_Z} =: t$. Our objective is to quantify the Monte-Carlo error

$$\delta r_{k,l} = \frac{\#\{i \in [\![1,p]\!] \mid \theta_i \in \Theta_{k,l}\}}{p} - \varpi(\Theta_{k,l}).$$

To that end, we fix $l \in [\![1,n]\!]$ and apply the standard Bernoulli Chernoff concentration inequality (additive form) to $X_i := \mathbb{1}(\theta_i \in \Theta_{k,l})$. By definition, $\mathbb{E}[X_i] = \varpi(\Theta_{k,l})$, hence by Chernoff

$$\mathbb{P}\left(\left|\frac{1}{p}\sum_{i=1}^{p} X_i - \varpi(\Theta_{k,l})\right| > t\right) \leq 2e^{-2pt^2}.$$

It follows that the inequality $p \geq \frac{\log(6n^2/\eta)}{2t^2}$ implies $|\delta r_{k,l}| \leq t$ with probability exceeding $1 - \frac{\eta}{3n^2}$. Substituting $t = \frac{\varepsilon}{4d n \bar{c}_Z}$ yields

$$(H_4):\ p \geq \frac{8d^2 n^2 \bar{c}_Z^2 \log(6n^2/\eta)}{\varepsilon^2}.$$

Using the same reasoning as in previous steps, under (H_4), the event $\{\forall l \in [\![1,n]\!],\ |\delta r_{k,l}| \leq \frac{\varepsilon}{4d n \bar{c}_Z}\}$ holds with probability exceeding $1 - \frac{\eta}{3n}$, which implies that our objective $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$ also holds with the same probability.

— *Step 6*: Condition for $\|\delta v_1\|_2 \leq \dfrac{\varepsilon}{4}$.

We have

$$\|\delta v_1\|_2 \leq \|\delta A^-\|_{\mathrm{op}} \left\|\sum_{l=1}^{n} r_{k,l} B_{k,l} z_l\right\|_2 \leq \frac{\|\delta A^-\|_{\mathrm{op}}}{d}\left(\|v\|_2 + \|\delta v_2\|_2 + \|\delta v_3\|_2 + \|\delta v_4\|_2\right).$$

In the following, we continue conditionally on the three events "$\|\delta v_i\|_2 \leq \frac{\varepsilon}{4}$", $i \in \{2,3,4\}$, under which:

$$\|\delta v_1\|_2 \leq \frac{\|\delta A^-\|_{\mathrm{op}}}{d}\left(\|v\|_2 + \frac{3\varepsilon}{4}\right).$$

We now dominate $\|v\|_2 = \left\|\sum_{l=1}^{n} S_{k,l} z_l\right\|_2$. Recall that the $(\Theta_{k,l})_{l \in [\![1,n]\!]}$ are disjoint, with $\bigcup_{l=1}^{n} \Theta_{k,l} = \mathbb{S}^{d-1}$, which implies $\sum_{l=1}^{n} S_{k,l} = d\int_{\mathbb{S}^{d-1}} \theta\theta^T \mathrm{d}\varpi(\theta) = I$. Since the $S_{k,l}$ are symmetric semi-definite, the previous equation provides $\|S_{k,l}\|_{\mathrm{op}} \leq 1$, which in turn yields $\|v\|_2 \leq n\bar{c}_Z$. Assuming $\varepsilon \leq \frac{4}{3} n \bar{c}_Z$, we get finally $\|\delta v_1\|_2 \leq \|\delta A^-\|_{\mathrm{op}} \frac{2n\bar{c}_Z}{d}$.

It is sufficient to find a condition under which $\|\delta A^-\|_{\mathrm{op}} \leq \frac{d\varepsilon}{8n\bar{c}_Z} =: t$. We cannot apply Lemma A.2 directly since δA^- has an inverse operation. First, $\|\delta A^-\|_{\mathrm{op}} = \|A^{-1} - dI\|_{\mathrm{op}} = \left\|d(I - d\delta A)^{-1} - dI\right\|_{\mathrm{op}}$, with $\delta A := I/d - A$. Then, assuming

$(H_{\delta A}) : d\|\delta A\|_{\mathrm{op}} < 1$, we use a Neumann series for the inverse:

$$\|\delta A^-\|_{\mathrm{op}} = \left\|\sum_{k=1}^{+\infty}(d\delta A)^k\right\|_{\mathrm{op}} \leq \sum_{k=1}^{+\infty}(d\|\delta A\|_{\mathrm{op}})^k,$$

and finally $\|\delta A^-\|_{\mathrm{op}} \leq \dfrac{d^2\|\delta A\|_{\mathrm{op}}}{1-d\|\delta A\|_{\mathrm{op}}}$. Consider $f := \left\{ \begin{array}{ccc} [0,\frac{1}{d}[& \longrightarrow & [0,+\infty[\\ u & \longmapsto & \frac{d^2u}{1-du} \end{array} \right.$.

The function f is bijective and increasing, with $f^{-1} = \left\{ \begin{array}{ccc} [0,+\infty[& \longrightarrow & [0,\frac{1}{d}[\\ v & \longmapsto & \frac{v}{d(d+v)} \end{array} \right.$.

This analysis yields under $(H_{\delta A})$, $\|\delta A^-\|_{\mathrm{op}} \leq t \Longleftarrow \|\delta A\|_{\mathrm{op}} \leq \dfrac{t}{d(d+t)}$.

Conveniently, by Lemma A.2, $\|\delta A\|_{\mathrm{op}} \leq s$ with probability $1-\eta/3$ if $p \geq \frac{8\log(3d/\eta)}{s^2}$. We can apply this to

$$\frac{t}{d(d+t)} = \frac{\varepsilon}{8dn\bar{c}_Z(1+\frac{\varepsilon}{8n\bar{c}_Z})},$$

but in order to simplify the expression, we apply it to

$$s := \frac{3\varepsilon}{28dn\bar{c}_Z} \leq \frac{t}{d(d+t)},$$

where the inequality holds thanks to $\varepsilon \leq \frac{4}{3}n\bar{c}_Z$.

Now we must quantify the assumption $(H_{\delta A}) : \|\delta A\|_{\mathrm{op}} < 1/d$. Notice that $s \leq 1/d$ and thus the event $\|\delta A\|_{\mathrm{op}} < s$ is contained in the event $\|\delta A\|_{\mathrm{op}} < 1/d$, hence it is sufficient to satisfy (H_1), which we write (after upper-bounding $8\times 28^2/9 \leq 697$):

$$(H_1): \quad p \geq \frac{697d^2n^2\bar{c}_Z^2\log(3d/\eta)}{\varepsilon^2}.$$

To summarise, under (H_1), we have $\|\delta A\|_{\mathrm{op}} \leq s$ with probability exceeding $1-\eta/3$. Conditionally to the events "$\|\delta A\|_{\mathrm{op}} \leq s$", "$\|\delta v_i\|_2 \leq \frac{\varepsilon}{4}$", $i \in \{2,3,4\}$, this step shows $\|\delta v_1\|_2 \leq \frac{\varepsilon}{4}$.

— *Step 7*: Wrapping up.

We now work under the conditions (H_i), $i \in \{1,2,3,4\}$. By Step 1,

$$\|y_k^* - v_k\|_2 \leq \|\delta v_1^k\|_2 + \|\delta v_2^k\|_2 + \|\delta v_3^k\|_2 + \|\delta v_4^k\|_2,$$

where we restore the omitted k indices. By Step 3, with probability exceeding $1-\eta/(3n)$, we have $\|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4}$, thus with probability $1-\eta/3$ we have $\forall k \in [\![1,n]\!]$, $\|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4}$. By Step 4, we have almost-surely $\forall k \in [\![1,n]\!]$, $\|\delta v_3^k\|_2 \leq \frac{\varepsilon}{4}$. By Step 5, with probability $1-\eta/3$, $\|\delta A\|_{\mathrm{op}} \leq s$. Putting this together yields that with probability $1-\eta$, we have:

$$\forall k \in [\![1,n]\!],\ \|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4},\ \|\delta v_3^k\|_2 \leq \frac{\varepsilon}{4},\ \|\delta v_4^k\|_2 \leq \frac{\varepsilon}{4} \text{ and } \|\delta A\|_{\mathrm{op}} \leq s.$$

Finally, Step 5 shows that conditionally to the events above, $\|\delta v_1^k\|_2 \leq \dfrac{\varepsilon}{4}$ almost-surely. Thus with probability exceeding $1-\eta$, $\forall k \in [\![1,n]\!], \|y_k^* - v_k\|_2 \leq \varepsilon$. Since $v_k = \sum_{l=1}^n S_{k,l}z_l$, with probability over $1-\eta$: $\forall k \in [\![1,n]\!]$, $\left\|y_k^* - \sum_{l=1}^n S_{k,l}z_l\right\|_2 \leq \varepsilon$. □

In order to get the summarised result from Section 3.2.3, we simplify the conditions as follows.

Corollary A.1 (Simplified conditions for Theorem A.2)**.** *With the notations of Theorem A.2, the condition:*
(A.15)

$$(H_p): \quad p \geq \left(\frac{4096 d^3 n \bar{c}_Z^3 \log(3dn^2/\eta)}{\varepsilon^3}\right) \vee \left(\frac{697 d^2 n^2 \bar{c}_Z^2 \log(3d/\eta)}{\varepsilon^2}\right) \vee \left(\frac{8 d^2 n^2 \bar{c}_Z^2 \log(6n^2/\eta)}{\varepsilon^2}\right)$$

implies (H_q) *and* $(H_i)_{i\in\{1,2,3,4\}}$*, and thus is sufficient in order to have (3.3).*

Proof. The second and third terms of (A.15) correspond to (H_1) and (H_4) respectively. Then, using $n^+, n^- \leq n$, we have

$$(H_2) \Longleftarrow \overline{q} \geq \frac{512 d^2 \bar{c}_Z^2 \log(3dn^2\eta)}{\varepsilon^2},$$

$$(H_3) \Longleftarrow \underline{q} \leq \frac{\varepsilon}{8dn\bar{c}_Z}\, p.$$

Let $q := \dfrac{512 d^2 \bar{c}_Z^2 \log(3dn^2/\eta)}{\varepsilon^2}$; $\overline{q} = \underline{q} = q$. (H_q) and (H_2) are automatically satisfied by this choice. For q to satisfy (H_3), it is sufficient to have

$$\frac{512 d^2 \bar{c}_Z^2 \log(3dn^2/\eta)}{\varepsilon^2} \leq \frac{\varepsilon}{8dn\bar{c}_Z}\, p, \text{ i.e. } p \geq \frac{4096 d^3 n \bar{c}_Z^3 \log(3dn^2/\eta)}{\varepsilon^3}$$

□

A.5. **Closed-form expression for Block-Coordinate Descent.** In Algorithm 1, we mention in line 4 the minimisation $\min_Y J(\pi, Y)$, where

$$J := \begin{cases} \mathbb{U}^p \times \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ (\pi^{(1)}, \cdots, \pi^{(p)}), Y & \longmapsto & \dfrac{1}{p}\displaystyle\sum_{i=1}^{p}\sum_{k=1}^{n}\sum_{l=1}^{n}(\theta_i^T y_k - \theta_i^T z_l)^2 \pi_{k,l}^{(i)} \end{cases},$$

and claim that it can in fact be done explicitly. We provide the formula below, which stems from a straightforward quadratic minimisation: let $Y^* = ((y_1^*)^T, \cdots, (y_n^*)^T)^T = \operatorname{argmin}_Y J(\pi, Y)$, we obtain

$$\forall k \in [\![1, n]\!],\ y_k^* = \left(\frac{1}{n}\sum_{i=1}^{p}\theta_i\theta_i^T\right)^{-1}\left(\sum_{i=1}^{p}\sum_{l=1}^{n}\pi_{k,l}^{(i)}\theta_i\theta_i^T z_l\right),$$

where we used the constraint $\pi \in \mathbb{U}^p$ which implies $\sum_l \pi_{k,l}^{(i)} = \frac{1}{n}$.

Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
Email address: eloi.tanguy@u-paris.fr

CMAP, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris
Email address: remi.flamary@polytechnique.edu

Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
Email address: julie.delon@u-paris.fr