Convergence Rates for Distribution Matching with Sliced Optimal Transport

Gauthier Thurin GTHURIN@MAIL.DI.ENS.FR
CNRS, ENS Paris, France

Claire Boyer
LMO, Université Paris-Saclay, Orsay, France ; Institut universitaire de France

Kimia Nadjahi
CNRS, ENS Paris, France

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the slice-matching scheme, an efficient iterative method for distribution matching based on sliced optimal transport. We investigate convergence to the target distribution and derive quantitative non-asymptotic rates. To this end, we establish Łojasiewicz-type inequalities for the Sliced-Wasserstein objective. A key challenge is to control along the trajectory the constants in these inequalities. We show that this becomes tractable for Gaussian distributions. Specifically, eigenvalues are controlled when matching along random orthonormal bases at each iteration. We complement our theory with numerical experiments and illustrate the predicted dependence on dimension and step-size, as well as the stabilizing effect of orthonormal-basis sampling.

Keywords: distribution matching, Sliced-Wasserstein distance, computational optimal transport, non-convex optimization, stochastic gradient descent

1. Introduction

Many problems in modern machine learning require comparing and matching probability distributions, as in generative modeling (Marzouk et al., 2016; Grenioux et al., 2023), density estimation (Wang and Marzouk, 2022; Irons et al., 2022) or domain adaptation (Courty et al., 2016). The goal is typically to transform a source distribution in order to match a more complex target distribution.

Distribution matching and optimal transport. Distribution matching can be naturally formalized through optimal transport (OT), which provides both a geometrically meaningful distance between probability measures and, when it exists, a transport map pushing a source distribution σ to a target distribution μ (Villani, 2008; Ambrosio and Savaré, 2007). OT-based methods have led to major theoretical and algorithmic advances across machine learning, image processing and scientific computing (Peyré et al., 2019; Santambrogio, 2015). However, computing OT maps is in general expensive both computationally and statistically (Hütter and Rigollet, 2021; Chewi et al., 2024).

Iterative approaches and measure interpolations. The high cost of OT has motivated alternative approaches that decompose the transport problem into simpler subproblems. A key idea is to build an interpolation between σ and μ through a sequence of elementary transformations, rather than estimating a single global transport map. This idea underlies many iterative correction schemes: although each step may only partially reduce the discrepancy between σ and μ, their composition is expected to gradually align them. Among all possible interpolations, the McCann interpolation (McCann, 1997) plays a distinguished theoretical role, as it corresponds to geodesics in Wasserstein space, but it is rarely tractable. A generic iterative sequence of measures that mimics McCann's interpolation can be constructed through

$$\widehat{\sigma}_{k+1} = \big((1-\gamma_k)\mathrm{Id} + \gamma_k \widehat{T}_k\big)_\sharp \widehat{\sigma}_k, \tag{1}$$

where $\widehat{T}_k$ is an approximate transport map from $\widehat{\sigma}_k$ to μ, $(\gamma_k)_k$ a sequence of step sizes. Here, $T_\sharp\sigma$ denotes the pushforward of σ by the function T: if $X \sim \sigma$, then $T(X) \sim T_\sharp\sigma$.

Different choices for $\widehat{T}_k$ have been proposed, such as entropy-regularized OT (Kassraie et al., 2024) and neural-network parameterizations in diffusion or flow-based models (Song et al., 2021; Albergo et al., 2025). In this work, we focus on sliced optimal transport, a computationally efficient alternative that leverages one-dimensional projections (Pitié et al., 2007; Rabin et al., 2011, 2012).

Sliced optimal transport and slice-matching maps. The Sliced-Wasserstein distance (SW) compares two distributions by projecting them onto one-dimensional subspaces and averaging the resulting Wasserstein distances (Rabin et al., 2011, 2012). Thanks to its scalability and simple implementation, SW has attracted growing interest in large-scale applications, including generative modeling (Deshpande et al., 2019; Wu et al., 2019; Liutkus et al., 2019; Kolouri et al., 2018; Dai and Seljak, 2021; Coeurdoux et al., 2022; Du et al., 2023). This empirical success has in turn motivated theoretical work on the geometry induced by sliced OT, sample complexity, and convergence properties of associated algorithms (Nadjahi et al., 2019, 2020; Manole et al., 2022; Tanguy, 2023; Tanguy et al., 2025; Li et al., 2023; Vauthier et al., 2025).

Although sliced OT does not directly provide transport maps or geodesics (Kitagawa and Takatsu, 2024; Park and Slepčev, 2025), several constructions have been proposed in this spirit (Liu et al., 2025; Mahey et al., 2023). In particular, *slice-matching maps* (Pitié et al., 2007; Li and Moosmüller, 2024) correspond to Wasserstein gradients of the SW functional (Li et al., 2023). For a direction $\theta \in \mathbb{S}^{d-1}$, let σ^θ and μ^θ denote the push-forwards of σ and μ by the projection $x \mapsto \langle x, \theta\rangle$. Denoting by $T_{\sigma^\theta}^{\mu^\theta} : \mathbb{R} \to \mathbb{R}$ the univariate optimal transport map from σ^θ to μ^θ, the associated slice-matching map is defined by

$$T_{\sigma,\theta}(x) = x + \big(T_{\sigma^\theta}^{\mu^\theta}(\theta^T x) - \theta^\top x\big)\theta\,. \tag{2}$$

Since the probability mass is transported along a single direction, $T_{\sigma,\theta}$ does not transport σ to μ. The *Iterative Distribution Transfer* (IDT) algorithm (Pitié et al., 2007) therefore constructs an iterative composition of slice-matching maps, corresponding to (1) with constant step sizes $\gamma_k = 1$. Using random directions θ at each iteration, this procedure is expected to gradually push σ to μ and has been successfully applied in practice.

Related works. The IDT algorithm (Pitié et al., 2007) was introduced before the Sliced-Wasserstein distance (Rabin et al., 2011, 2012) and was later interpreted as an iterative sliced OT procedure. Early works established convergence of the IDT iterates when the target is the standard Gaussian distribution and studied its continuous-time limit, often referred to as the *Sliced-Wasserstein flow* (Pitié

et al., 2007; Bonnotte, 2013). More recently, Cozzi and Santambrogio (2025) proved convergence of SW flows to the isotropic Gaussian. Relatively few results are available on the convergence of sliced OT procedures beyond the Gaussian setting. A more general analysis is conducted in Li et al. (2023), which reinterprets IDT as a stochastic gradient descent method (SGD) on SW and accounts for time discretization and randomness in the sampled directions. They prove asymptotic convergence of the discrete-time dynamics under strong assumptions, notably that the iterates remain in a compact set containing no other critical points than the target measure. In parallel, several works have studied SW as a loss between discrete measures and highlight the existence of nontrivial critical points, which motivate noisy or regularized variants of SGD (Tanguy et al., 2024, 2025; Vauthier et al., 2025).

Contributions. The main goal of this paper is to establish convergence rates for the slice-matching scheme (Li et al., 2023). Our approach is based on identifying Polyak–Łojasiewicz (PL) inequalities for the Sliced-Wasserstein objective, which bound the loss by the squared norm of its Wasserstein gradient. These inequalities imply quantitative convergence rates to the target distribution. The main technical challenge is that the associated constants depend on lower and upper bounds on the density of the iterates, which are difficult to control along the trajectory.

We address this difficulty within the class of elliptic distributions, for which slice-matching maps are linear. In this regime, controlling the density of the iterates amounts to controlling the eigenvalues of their covariance matrices. When the target distribution is isotropic, we show that these eigenvalues can be controlled in expectation, which in turn yields explicit convergence rates. Crucially, such spectral control holds from the very first iteration when the updates use random orthonormal bases of directions. This stands in contrast with the single-direction setting, where the lack of orthogonality leads to larger fluctuations in the covariance structure before stabilization.

Structure. Section 2 introduces the mathematical framework. Preliminary convergence results to critical points are discussed in Section 3. Section 4 presents our main results on Łojasiewicz- and PL-type inequalities and on the control of the associated constants. Numerical experiments are reported in Section 5, followed by a conclusion. Technical proofs are deferred to the appendices.

Notation. For any probability measure ν on $\mathbb{R}^d$, let $\mathrm{M}_2(\nu) = \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\nu(x)$ be its second moment. $\mathcal{P}_2(\mathbb{R}^d)$ refers to the set of measures with a finite second moment and $\mathcal{P}_{2,ac}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ is the set of absolutely continuous measures with respect to the Lebesgue measure. We denote the Euclidean norm and inner product on $\mathbb{R}^d$ by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$. For $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, we define $\mathbf{L}^2(\nu) = \{f : \mathbb{R}^d \to \mathbb{R}^d \ : \ \int_{\mathbb{R}^d} \|f(x)\|^2 \mathrm{d}\nu(x) \leq +\infty\}$, and for $f, g \in \mathbf{L}^2(\nu)$, $\langle f, g \rangle_\nu = \int_{\mathbb{R}^d} \langle f(x), g(x) \rangle \mathrm{d}\nu(x)$, $\|f\|_\nu = \sqrt{\langle f, f \rangle_\nu}$. Let $\mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ be the unit sphere in $\mathbb{R}^d$. For any $\theta \in \mathbb{S}^{d-1}$, $\pi_\theta : \mathbb{R}^d \to \mathbb{R}$ is the projection $\pi_\theta(x) = \langle x, \theta \rangle$. Finally, $\lambda_i(A)$ refers to the i-th smallest eigenvalue of a matrix A, with $\lambda_{\min}(A)$ the smallest and $\lambda_{\max}(A)$ the largest.

2. Background on the Slice-Matching Scheme

We begin by reviewing the definition of optimal transport and its properties for one-dimensional measures, which motivates slicing. Let T_σ^μ denote the OT map from σ to μ, defined as a minimizer in the Wasserstein distance: $W_2^2(\sigma, \mu) = \inf_{T : T_\sharp \sigma = \mu} \mathbb{E}_{X \sim \sigma} \big\| X - T(X) \big\|^2$. In dimension one, the optimal transport map admits a closed-form expression, $T_\sigma^\mu = F_\mu^{-1} \circ F_\sigma$, where F_ρ is the cumulative distribution function of $\rho \in \mathcal{P}_2(\mathbb{R})$. This motivates the definition of the Sliced-Wasserstein distance,

which averages one-dimensional Wasserstein distances over random projections:

$$SW_2^2(\sigma, \mu) = \int_{\mathbb{S}^{d-1}} W_2^2(\sigma^\theta, \mu^\theta)\, \mathrm{d}\mathcal{U}(\theta),$$

where $\sigma^\theta = (\pi_\theta)_\sharp \sigma$ and $\mu^\theta = (\pi_\theta)_\sharp \mu$, and $\mathcal{U}$ is the uniform distribution on $\mathbb{S}^{d-1}$.

Slice-matching maps and scheme. We now introduce the slice-matching construction that underlies the iterative scheme studied in this paper. Let $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ be a target probability measure, and let $P = [\theta_1, \dots, \theta_d] \in \mathbb{R}^{d\times d}$ be an orthonormal basis of $\mathbb{R}^d$. For any direction $\theta \in \mathbb{S}^{d-1}$, denote by $t_\theta = T_{\sigma^\theta}^{\mu^\theta}$ the one-dimensional optimal transport map pushing the projected measure σ^θ onto μ^θ.

Rather than transporting mass along a single direction, we simultaneously match d orthogonal one-dimensional projections. This leads to the definition of the (matrix-)slice-matching map

$$\forall x \in \mathbb{R}^d, \quad T_{\sigma,P}(x) = x + P \begin{bmatrix} t_{\theta_1}(\theta_1^\top x) - \theta_1^\top x \\ t_{\theta_2}(\theta_2^\top x) - \theta_2^\top x \\ \vdots \\ t_{\theta_d}(\theta_d^\top x) - \theta_d^\top x \end{bmatrix} = \sum_{i=1}^d t_{\theta_i}(\theta_i^\top x)\, \theta_i, \tag{3}$$

where the last equality follows from the fact that P is an orthonormal basis. Using several orthogonal directions at each iteration has been observed to significantly improve both stability and empirical performance (Pitié et al., 2007; Bonneel et al., 2015; Li et al., 2023). From a theoretical standpoint, matrix-slice-matching maps enjoy a moment-matching property (Li and Moosmüller, 2024, Proposition 3.6), which will play a central role in our analysis:

$$\mathbb{E}_{Y\sim(T_{\sigma,P})_\sharp\sigma}[Y] = \mathbb{E}_{Y\sim\mu}[Y], \qquad \mathrm{M}_2\big((T_{\sigma,P})_\sharp\sigma\big) = \mathrm{M}_2(\mu).$$

The *slice-matching scheme*, main focus of this paper, is defined as follows: Starting from an initial distribution $\sigma_0 = \sigma \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, the iterates are given by

$$\forall k \geq 0, \qquad \sigma_{k+1} = \big((1-\gamma_k)\mathrm{Id} + \gamma_k T_{\sigma_k, P_{k+1}}\big)_\sharp \sigma_k, \tag{4}$$

where $(P_k)_{k\geq 1}$ is an i.i.d. sequence of random orthonormal bases drawn according to the Haar measure on $O(d)$ (the set of $d \times d$ orthonormal matrices) and $(\gamma_k)_{k\geq 0}$ consist of positive step sizes satisfying the Robbins-Monro conditions

$$\sum_{k\geq 0} \gamma_k = +\infty, \qquad \sum_{k\geq 0} \gamma_k^2 < +\infty. \tag{5}$$

Stochastic gradient descent perspective. The slice-matching scheme admits a natural interpretation as a stochastic gradient descent procedure in the 2-Wasserstein space for a Sliced-Wasserstein loss (Li et al., 2023). Specifically, consider the variational problem

$$\min_{\sigma\in\mathcal{P}_2(\mathbb{R}^d)} \mathscr{F}(\sigma), \qquad \text{with} \qquad \mathscr{F}(\sigma) = \frac{d}{2} SW_2^2(\sigma, \mu). \tag{6}$$

For $P = [\theta_1, \dots, \theta_d]$ an orthonormal basis of $\mathbb{R}^d$, defining $\mathscr{F}(\sigma, P) = \frac{1}{2}\sum_{\ell=1}^d W_2^2(\sigma^{\theta_\ell}, \mu^{\theta_\ell})$, one has the decomposition

$$\mathscr{F}(\sigma) = \mathbb{E}_P\big[\mathscr{F}(\sigma, P)\big],$$

where the expectation is taken with respect to P[1]. Both $\mathscr{F}$ and $\mathscr{F}(\cdot, P)$ depend on the target measure μ, a dependence that we omit in the notation for simplicity. The Wasserstein gradient of the random functional $\mathscr{F}(\cdot, P)$ is given by

$$\nabla_{W_2}\mathscr{F}(\sigma, P) = \mathrm{Id} - T_{\sigma,P},$$

and provides an unbiased estimator of the full Wasserstein gradient: $\mathbb{E}_P\big[\nabla_{W_2}\mathscr{F}(\sigma, P)\big] = \nabla_{W_2}\mathscr{F}(\sigma)$, see Rabin et al. (2011); Bonnotte (2013); Li et al. (2023); Cozzi and Santambrogio (2025) and Proposition 6 (Appendix B). As a consequence, the slice-matching iteration (4) can be rewritten as a stochastic gradient descent update in Wasserstein space. For any $k \geq 0$,

$$\sigma_{k+1} = \big(\mathrm{Id} - \gamma_k(\mathrm{Id} - T_{\sigma_k,P_{k+1}})\big)_\sharp \sigma_k = \big(\mathrm{Id} - \gamma_k \nabla_{W_2}\mathscr{F}(\sigma_k, P_{k+1})\big)_\sharp \sigma_k. \tag{7}$$

For completeness, Appendix A recalls basic notions of differentiation in Wasserstein space.

Bounded gradients. Cozzi and Santambrogio (2025) show that second-order moments are bounded along the Sliced-Wasserstein flow. In our discrete time setting that incorporates stochastic choices of directions P_{k+1}, we can show that the same holds as a result of the aforementioned moment-matching property of slice-matching maps (see Proposition 7, Appendix B). Combining this with $\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 \leq 2\mathscr{F}(\sigma)$ (by Jensen's inequality; see Proposition 6, Appendix B), one has

$$\forall k \geq 0, \quad \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k}^2 \leq 2\mathscr{F}(\sigma_k) \leq 4\mathrm{M}_2(\mu)\,.$$

Smoothness and non-convexity. A key property for SGD is the *smoothness* of the objective function. It is shown in Vauthier et al. (2025) (and recalled in Appendix B.2) that $\mathscr{F}$ is 1-smooth in $\mathcal{P}_2(\mathbb{R}^d)$ endowed with W_2: for any $\sigma_1, \sigma_2 \in \mathcal{P}_2(\mathbb{R}^d)$ such that the OT map $T_{\sigma_1}^{\sigma_2}$ exists,

$$\mathscr{F}(\sigma_2) \leq \mathscr{F}(\sigma_1) + \langle \nabla\mathscr{F}(\sigma_1), T_{\sigma_1}^{\sigma_2} - \mathrm{Id}\,\rangle_{\sigma_1} + \frac{1}{2}W_2^2(\sigma_1, \sigma_2). \tag{8}$$

Smoothness alone, however, is not sufficient to guarantee almost-sure convergence towards μ. In Wasserstein spaces, convergence rates typically rely on *geodesic convexity* (Ambrosio and Savaré, 2007), which $\mathscr{F}$ does not satisfy in general (Vauthier et al., 2025). Nevertheless, convergence is observed in practice (Pitié et al., 2007; Rabin et al., 2011), which suggests that the optimization landscape remains highly structured, as studied in the next section.

3. Preliminary Analysis: Convergence to Critical Points

We recall convergence results from Li et al. (2023) and derive new results about averages of gradient norms with standard proofs that use the smoothness property.

Descent lemma. The following lemma is a key recursion inequality that serves as a standard descent condition in stochastic optimization. The proof follows by the smoothness property (8) and direct computations, in the same fashion as for optimization over Euclidean spaces.

1. This equality follows from the invariance of the Haar measure, which ensures that the marginal distribution of each direction θ_ℓ is uniform on $\mathbb{S}^{d-1}$, even though the directions are not independent.

Lemma 1 (Li et al. (2023), Lemma A.1) *Let $(\sigma_k)_{k\geq 1}$ be the iterates generated by the slice-matching scheme (4). Then, for any $k \geq 0$,*

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})|\mathcal{A}_k] \leq (1+\gamma_k^2)\mathscr{F}(\sigma_k) - \gamma_k \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}\,, \tag{9}$$

where $\mathcal{A}_k$ is the σ-field generated by $(P_1, \dots, P_k)$.

Given recursion (9) and step-sizes assumptions (5), a direct application of Robbins-Siegmund theorem (Robbins and Siegmund, 1971) implies that $(\mathscr{F}(\sigma_k))_{k\geq 0}$ converges almost surely to a finite random variable, and that

$$\sum_{k\geq 1} \gamma_k \big\|\nabla_{W_2}\mathscr{F}(\sigma_k)\big\|^2_{\sigma_k} < +\infty \quad a.s. \tag{10}$$

An immediate byproduct is that a subsequence of $(\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k})_{k\geq 1}$ converges almost surely to 0, or equivalently $\liminf_{k\to+\infty} \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k} = 0$. Besides, $(\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k})_{k\geq 1}$ converges almost surely to 0 if the sequence $(\sigma_k)_{k\geq 1}$ remains in a compact subset of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ (Li et al., 2023, Theorem 2). This holds true for instance if σ_0 and μ are continuous and compactly supported, or under finite third-order moments (Li et al., 2023, Remark 9). Under the additional assumption that $\nabla\mathscr{F}(\sigma) = 0 \iff \sigma = \mu$, the limit of σ_k must be μ almost surely. To the best of our knowledge, the only known sufficient condition for this equivalence is that densities are strictly positive on their compact support (Bonnotte, 2013, Lemma 5.7.2).

Convergence Guarantees to Critical Points. The next proposition establishes convergence toward a critical point using standard arguments, up to a random reshuffling of the indices (Ghadimi and Lan, 2013). This result is weaker than the almost sure convergence $\sigma_k \overset{a.s.}{\to} \mu$ from Li et al. (2023, Theorem 2), but it has the benefit of requiring no additional assumptions than the ones of Lemma 1. Here, this means absolute continuity for σ and μ, although smoothness (8) holds in fact in the more difficult setting of Vauthier et al. (2025) where (σ_k) are discrete. In this case, the next two propositions could be extended.

Proposition 1 *For any $K \in \mathbb{N}$, let $i(K)$ be a random index such that $\forall k \in \{1, \dots, K\}$, $\mathbb{P}(i(K) = k) = 1/K$. Then, $(\|\nabla\mathscr{F}(\sigma_{i(K)})\|^2_{\sigma_{i(K)}})_{K\geq 0}$ converges in probability towards 0,* i.e.*,*

$$\forall \epsilon > 0, \quad \lim_{K\to+\infty} \mathbb{P}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|^2_{\sigma_{i(K)}} > \epsilon\big) = 0\,.$$

Turning to convergence rates, assuming smoothness and boundedness of the iterates only yields the following result, which concerns a weighted average of the gradients.

Proposition 2 *For a number K of iterations, define the weigths $\omega_j = \gamma_j / \sum_{k=1}^K \gamma_k$, for any $0 \leq j \leq K$, where $(\gamma_j)_j$ are the chosen learning rates. Then,*

$$\sum_{k=0}^K \omega_k \mathbb{E}[\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}] \leq \frac{\mathscr{F}(\sigma_0) + 4\mathrm{M}_2(\mu)\sum_{k=0}^K \gamma_k^2}{\sum_{k=0}^K \gamma_k}\,. \tag{11}$$

When choosing $\gamma_k = 1/(k+1)^\alpha$ for $1/2 < \alpha < 1$, considering that the numerator is bounded by a constant, Proposition 2 yields a rate of order $K^{\alpha-1}$, since $\sum_{k=0}^K \gamma_k \geq \frac{1}{1-\alpha}(K^{1-\alpha} - 1)$. We also

emphasize that the bound (11) would tend to zero for a constant step-size $\gamma_k = 1/\sqrt{K+1}$ given a finite time horizon K (as in Ghadimi and Lan, 2013; Khaled and Richtárik, 2023).

These propositions complement the related work by Vauthier et al. (2025) that also study convergence towards critical points. Their setting is different in that they consider a discrete source σ, a continuous target μ, a constant learning rate and their gradients are theoretically computed from all directions $\theta \in \mathbb{S}^{d-1}$, as opposed to our stochastic gradients along finitely many directions.

The convergence results obtained so far are standard for stochastic optimization of smooth losses with bounded gradients (Bottou et al., 2018; Dossal et al., 2024). For completeness, proofs are provided in Appendix B.4. In the remainder of this paper, we will assume appropriate continuity conditions, allowing us to strengthen and extend the preceding results.

In particular, our Łojasiewicz inequalities imply that the assumptions of Li et al. (2023, Theorem 2) hold for Gaussian measures. This readily gives almost-sure convergence, as stated hereafter and proved in Appendix E.1.

Proposition 3 *Let $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \Lambda)$, with $\Sigma, \Lambda \in \mathbb{R}^{d\times d}$ strictly positive definite. Let (γ_k) satisfy the Robbins-Monro conditions (5). Then, $\lim_{k\to+\infty} \mathcal{F}(\sigma_k) = 0$ almost surely.*

The almost-sure convergence of the objective can be converted into convergence in W_2, using the compactness of the iterates and the metric properties of SW_2, as in the proof of (Cozzi and Santambrogio, 2025, Corollary 4.3).

Corollary 1 *Under the assumptions of Proposition 3, $\lim_{k\to+\infty} W_2(\sigma_k, \mu) = 0$ almost surely.*

4. Convergence Analysis under Łojasiewicz Inequalities

This section is devoted to the derivation of quantitative convergence rates for the slice-matching scheme. Our main result concerns Gaussian source and target measures.

4.1. Main result: convergence analysis for Gaussian measures

Theorem 2 *Assume $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \mathbf{I}_d)$, where $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric positive definite. Let $\gamma_k = 1/(k+1)^\alpha$. For $2/3 < \alpha < 1$, it exists $C > 0$ such that, for all $k \geq 1$,*

$$\mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{C}{k^{2\alpha-1}}.$$

For $0 < \alpha < 2/3$, it exists $C > 0$ such that, for all $k \geq 1$ and for all $0 < \epsilon < \min(\alpha, 1-\alpha)$,

$$\mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{C}{k^{1-\alpha-\epsilon}}.$$

The complete proof is deferred to Appendix E. The remainder of this section presents the main ingredients and is organized as follows. We first introduce a general framework showing how convergence rates follow from a *random Polyak–Łojasiewicz (PL) inequality* along the trajectory. We then discuss how such inequalities can be established in a *static* fashion under density bounds, and why propagating these bounds is difficult in general. Finally, we show that the Gaussian structure allows one to control the corresponding PL constants through spectral estimates on covariance matrices, which leads to Theorem 2.

4.2. Step 1: From (random) PL inequalities to rates

Our starting point is a gradient-variance decomposition (see Appendix B.1, Proposition 6) which isolates Łojasiewicz-type inequalities as the key ingredient. Denoting $\overline{T}_\sigma = \mathbb{E}_P[T_{\sigma,P}]$, one has

$$2\mathscr{F}(\sigma) = \|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 + \mathbb{E}_P\big[\|\overline{T}_\sigma - T_{\sigma,P}\|_\sigma^2\big]\,. \tag{12}$$

If the variance term is controlled by the squared Wasserstein gradient norm, *i.e.*, if there exists $s > 0$ such that

$$\mathbb{E}_P\big[\|\overline{T}_\sigma - T_{\sigma,P}\|_\sigma^2\big] \le s^2\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2,$$

then (12) yields a Polyak–Łojasiewicz inequality

$$\mathscr{F}(\sigma) \le B\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2, \qquad B = \tfrac{1+s^2}{2},$$

a standard condition to prove convergence rates in nonconvex optimization (*e.g.* Garrigos and Gower, 2023). This motivates the search for PL inequalities that hold *along the iterates* $(\sigma_k)_{k\ge0}$ with constants that can be controlled. We formalize this requirement through the following random Łojasiewicz-type condition (Kurdyka et al., 2000; Attouch et al., 2010).

Assumption A *For some $\tau \in \{1,2\}$ and any $k \ge 1$, $\mathscr{F}(\sigma_k)^\tau \le B_k\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k}^2$ with $(B_k)_{k\ge1}$ a sequence of positive random variables s.t. $\sup_{k\ge1}\mathbb{E}[B_k^p] \le c_p$ with $c_p \in (0,+\infty)$ for all $p \in \mathbb{N}^*$.*

By combining such inequalities along the trajectory with the descent recursion for $\mathscr{F}(\sigma_k)$ (Lemma 1), we obtain the following rates.

Theorem 3 *Consider Assumption A with $\tau = 1$. Choose the step sequence as $\gamma_k = 1/(k+1)^\alpha$.*

(i) If $0 < \alpha < 2/3$, then, for any $k \ge 1$, $\mathbb{E}[\mathscr{F}(\sigma_k)] \lesssim k^{-(1-\alpha-\epsilon)}$ for all $0 < \epsilon < \min(\alpha, 1-\alpha)$.

(ii) If $2/3 < \alpha < 1$, then, for any $k \ge 1$, $\mathbb{E}[\mathscr{F}(\sigma_k)] \lesssim k^{-(2\alpha-1)}$.

Alternatively, consider Assumption A with $\tau = 2$. For $p \ge 2\alpha/(2-3\alpha)$, let $\gamma = (2\mathrm{M}_2(\mu)\sqrt{c_p})^{3/2}$. Let $\gamma_k = 1/(k+\gamma)^\alpha$ with $1/2 < \alpha < 2/3$. Then, for any $k \ge 1$, $\mathbb{E}[\mathscr{F}(\sigma_k)] \lesssim 1/(k+\gamma)^{2\alpha-1}$.

Only finitely many moments of B_k are required for the analysis. More precisely, the proof requires $\sup_{k\ge1}\mathbb{E}[B_k^p] < \infty$ for some $p > 4\alpha/(1-\alpha)$ when $\tau = 1$, and for some $p \ge 2\alpha/(2-3\alpha)$ when $\tau = 2$. For simplicity of exposition, Assumption A is stated with uniform bounds for all $p \in \mathbb{N}^*$.

Beyond the slice-matching setting, the proof strategy applies more generally to optimization schemes with smooth objectives, whose gradients are bounded and that satisfy Assumption A. The argument follows a standard template: one first derives a descent recursion, and then applies an appropriate variant of Chung's lemma (Chung, 1954; Jiang et al., 2024). The main additional difficulty here is that the PL constant B_k is random. We address this by working on events of the form $\{B_k \le g_k^{-1}\}$ where $g_k \to 0$ is chosen so that these events eventually occur almost surely. Similar arguments appear in Godichon-Baggioni (2019, Theorem 4.2) and Bercu and Bigot (2021, Theorem 3.6) to leverage local strong convexity. The main remaining difficulty is therefore to verify Assumption A for the slice-matching iterates.

4.3. Step 2: Static PL inequalities and bounded densities

In this section, we show that Łojasiewicz-type inequalities can be established in a *static* manner, *i.e.*, for fixed measures with uniformly bounded densities.

Gradient domination for bounded densities. For notational simplicity, we identify any $\sigma \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ with its density. Given a reference measure $\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, we consider the convenient setting of measures with uniformly bounded densities

$$\mathcal{P}_{\nu,m,M}(\mathbb{R}^d) = \{\sigma \in \mathcal{P}(\mathbb{R}^d) : m\nu \leq \sigma \leq M\nu\}\,, \tag{13}$$

for which the following gradient-domination inequality can be obtained.

Proposition 4 *Assume that $\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ satisfies a Poincaré inequality with constant $C_\nu > 0$, i.e., for any $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|\nabla f\|_\nu^2 < +\infty$, $\mathrm{Var}_\nu(f) \triangleq \|f - \mathbb{E}_\nu[f]\|_\nu^2 \leq C_\nu \|\nabla f\|_\nu^2$. Then, if $\mu \in \mathcal{P}_{\nu,m,M}(\mathbb{R}^d)$, for any $\sigma \in \mathcal{P}_{\nu,m,M}(\mathbb{R}^d)$,*

$$\mathscr{F}(\sigma) \leq 2C_\nu \frac{M}{m} \left\|\nabla_{W_2}\mathscr{F}(\sigma)\right\|_\sigma .$$

The proof follows arguments similar to Chizat et al. (2025, Lemma 3.3). Note that, combined with $\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 \leq 2\mathscr{F}(\sigma)$, we obtain the two-sided estimate

$$\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2/2 \leq \mathscr{F}(\sigma) \leq 2C_\nu \frac{M}{m} \left\|\nabla_{W_2}\mathscr{F}(\sigma)\right\|_\sigma .$$

In particular, $\nabla_{W_2}\mathscr{F}(\sigma) = 0$ if and only if $\mathscr{F}(\sigma) = 0$, *i.e.*, $\sigma = \mu$. We therefore retrieve a characterization of critical points by Bonnotte (2013, Lemma 5.7.2), where compactness of the support is no longer required.

PL inequality for Gaussians. We now turn to the Gaussian setting, in which PL inequalities can be established. We consider the class

$$\mathcal{G}_{m,M} = \{\rho_\Sigma : \Sigma \in \mathrm{S}_{++}^d,\ m\mathbf{I}_d \preceq \Sigma \preceq M\mathbf{I}_d\}, \tag{14}$$

where $\rho_\Sigma = \mathcal{N}(0, \Sigma)$, and S_{++}^d is the set of positive definite $d \times d$ matrices. The notation $\preceq$ refers to the Loewner partial order: for two symmetric matrices (A, B), $A \preceq B$ if and only if $B - A$ is positive semi-definite. Therefore, $\mathcal{G}_{m,M}$ corresponds to Gaussian measures with uniformly bounded covariance eigenvalues.

Proposition 5 (PL inequality on $\mathcal{G}_{m,M}$) *Let $\sigma = \rho_\Sigma$ and $\mu = \rho_\Lambda$ such that Σ, Λ are simultaneously diagonalizable by an orthogonal matrix (i.e., co-diagonalizable). Assume $\rho_\Sigma, \rho_\Lambda \in \mathcal{G}_{m,M}$. Let $C_d = d(d+2)M/m$. Then,*

$$\mathscr{F}(\sigma) \ \leq\ \frac{C_d}{2}\Big(1 + \frac{M}{m}\Big)\ \left\|\nabla_{W_2}\mathscr{F}(\sigma)\right\|_\sigma^2 . \tag{15}$$

Proposition 5 is proved by adapting Chewi et al. (2020, Theorem 19), which yields an intermediate inequality relating $\mathscr{F}(\sigma)$ and $\|\nabla_{W_2}\mathscr{F}(\sigma)\|$ for $\sigma, \mu \in \mathcal{G}_{m,M}$ (see Appendix C.2). We then refine it into a PL inequality by proving that, for co-diagonalizable covariances,

$$W_2^2(\rho_\Sigma, \rho_\Lambda) \leq C_d\, SW_2^2(\rho_\Sigma, \rho_\Lambda). \tag{16}$$

To our knowledge, this is the first comparison between W_2 and SW_2 with polynomial dimension dependence, instead of exponential dependence obtained in general settings, *e.g.*, Bonnotte (2013, Theorem 5.1.5) and Carlier et al. (2025). This result may be of independent interest for other research problems involving Gaussian distributions and the Bures-Wasserstein metric.

From static inequalities to iterate stability. To use Proposition 4 (or Proposition 5) in a convergence analysis, one must ensure that the iterates $(\sigma_k)_{k\geq 0}$ remain in $\mathcal{P}_{\nu,m,M}(\mathbb{R}^d)$ (or $\mathcal{G}_{m,M}$) with constants m, M uniform in k. However, if σ_k satisfies such bounds, propagating them to σ_{k+1} is challenging. Indeed, since $\sigma_{k+1} = S_{k\#}\sigma_k$, the change-of-variables formula yields

$$\sigma_{k+1}(S_k(x)) = \frac{\sigma_k(x)}{\det \text{Jac}\,[S_k](x)}. \tag{17}$$

Thus, propagating density bounds reduces to controlling $\det \text{Jac}\,[S_k]$, which typically requires strong regularity estimates on S_k; see *e.g.*, Caffarelli (1992, 2000); Bobkov and Ledoux (2019); Park and Slepčev (2025). One possible way to circumvent this difficulty in general settings is to introduce diffusion through entropic regularization (Chizat et al., 2025), but this leads to a different class of distribution-matching algorithms (Liutkus et al., 2019) and falls outside the scope of the present work. This observation motivates restricting attention to settings, such as the Gaussian case, where the relevant constants can instead be controlled through an alternative, more tractable mechanism.

4.4. Step 3: Propagating PL constants along the trajectory in the Gaussian case

To apply Theorem 3, it remains to verify Assumption A along the slice-matching trajectory, in the Gaussian setting.

Slice-matching on the Bures-Wasserstein manifold. We begin by making the slice-matching updates explicit when matching two Gaussians. Let $\mu = \rho_\Lambda$ and for a fixed $k \in \mathbb{N}$, $\sigma_k = \rho_{\Sigma_k}$. Then, for any $\theta \in \mathbb{S}^{d-1}$, the one-dimensional projections satisfy $\sigma_k^\theta = \mathcal{N}(0, \theta^\top \Sigma \theta)$, $\mu^\theta = \mathcal{N}(0, \theta^\top \Lambda \theta)$, and the corresponding optimal transport map between these marginals is linear and given by

$$T_{\sigma_k^\theta}^{\mu^\theta}(s) = \tau_\theta s, \quad \text{with} \quad \tau_\theta = \sqrt{\theta^\top \Lambda \theta / \theta^\top \Sigma \theta}.$$

For $P_{k+1} = [\theta_1, \cdots, \theta_d]$, define the diagonal matrix $D_k = \text{diag}(\tau_{\theta_1}, \ldots, \tau_{\theta_d})$. The resulting slice-matching map $T_{\sigma_k, P_{k+1}}$ is also linear: $\forall x \in \mathbb{R}^d$, $T_{\sigma_k,P_{k+1}}(x) = P_{k+1} D_k P_{k+1}^\top x$. As a consequence, the iterates remain Gaussian (Altschuler et al., 2021), *i.e.*, $\sigma_k = \rho_{\Sigma_k}$, with covariance matrices evolving according to the following recursion

$$\Sigma_{k+1} = A_k \, \Sigma_k \, A_k^\top, \qquad A_k = (1-\gamma_k)\mathbf{I}_d + \gamma_k P_{k+1} D_k P_{k+1}^\top. \tag{18}$$

Remark 4 (Centered Gaussians) *If $\gamma_1 = 1$, the first iteration enforces equality of the means of σ_1 and μ due to the moment matching property of slice-matching maps (Li and Moosmüller, 2024, Proposition 3.6). Therefore, we may assume without loss of generality that σ and μ are centered.*

Remark 5 (Elliptically contoured distributions) *All results of this section extend beyond the Gaussian case to elliptically contoured distributions. The key structural property used throughout is the linearity of OT maps, which also holds in this broader class (Gelbrich, 1990, Theorem 2.1).*

Control of PL constants along the trajectory. The convergence analysis relies on PL inequalities whose constants depend inversely on the smallest eigenvalue of the covariance matrices Σ_k. Therefore, obtaining quantitative convergence rates requires uniform (in k) control of $1/\lambda_{\min}(\Sigma_k)$, in expectation and with finite moments. We thus proceed in three steps:

(a) "Static" Łojasiewicz inequalities with random constants. Under the trace bound $\mathrm{Tr}(\Sigma_k) \le \mathrm{Tr}(\Lambda)$ (Proposition 7), one has $\lambda_{\min}(\Sigma_k) \le \lambda_{\max}(\Lambda)$. Consequently, Propositions 5 and 8 yield, for $\tau \in \{1, 2\}$,

$$\mathscr{F}(\sigma_k)^\tau \;\le\; B_k \left\|\nabla_{W_2}\mathscr{F}(\sigma_k)\right\|_{\sigma_k}^2, \quad B_k \;\lesssim\; \frac{1}{\lambda_{\min}(\Sigma_k)}. \tag{19}$$

Thus, the PL constant along the trajectory is random and may deteriorate if $\lambda_{\min}(\Sigma_k)$ becomes small, which motivates a quantitative control of this quantity.

(b) Recursion on $\lambda_{\min}(\Sigma_{k+1})$. We exploit the explicit covariance update (18). We show that for any $k \ge 0$, there exists a direction θ_i among the columns of P_{k+1} such that (Proposition 10)

$$\sqrt{\lambda_{\min}(\Sigma_{k+1})} \;\ge\; \sqrt{\lambda_{\min}(\Sigma_k)}\big(1 - \gamma_k + \gamma_k \tau_{\theta_i}\big), \quad \tau_{\theta_i} = \sqrt{\frac{\theta_i^\top \Lambda \theta_i}{\theta_i^\top \Sigma_k \theta_i}}. \tag{20}$$

Since $\Sigma_0 \succ 0$ and $\Lambda \succ 0$, it holds by induction that $\lambda_{\min}(\Sigma_k) > 0$ for all finite k. Hence, the PL inequality in Proposition 5 is well-defined along the trajectory.

(c) Moment control of $1/\lambda_{\min}(\Sigma_k)$. We now leverage the recursion (20) to bound $1/\lambda_{\min}(\Sigma_k)$ in expectation. A sufficient condition is provided by Proposition 12: for some $p \ge 1$,

$$\mathbb{E}\left[\sum_{k\ge 0} \gamma_k \, \mathbb{E}_\theta\left[\left(\frac{\theta^\top \Sigma_k \theta}{\theta^\top \Lambda \theta}\right)^p - 1\right]\right] < \infty. \tag{21}$$

We are able to verify (21) in the isotropic target case $\Lambda = \mathbf{I}_d$, although our numerical experiments suggest that (21) is verified for more general target covariances. More precisely, for any $p \in \mathbb{N}^*$, $\sup_{k\ge 1} \mathbb{E}[\lambda_{\min}(\Sigma_k)^{-p}] < \infty$ when $\Lambda = \mathbf{I}_d$ (Proposition 11).

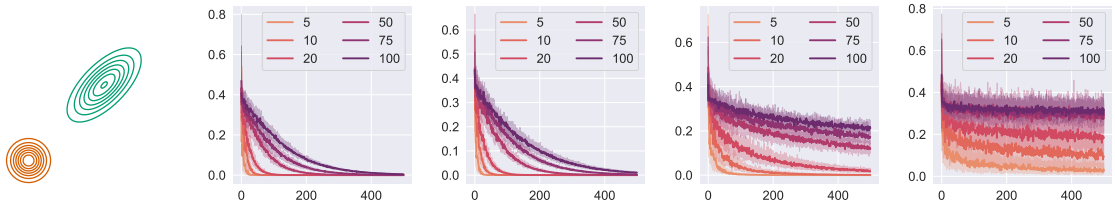
Combining the PL inequality (19) with the above moment bounds shows that Assumption A holds along the Gaussian slice-matching trajectory. Applying Theorem 3 then yields the convergence rate stated in Theorem 2.

5. Numerical Experiments

5.1. Matching Gaussians

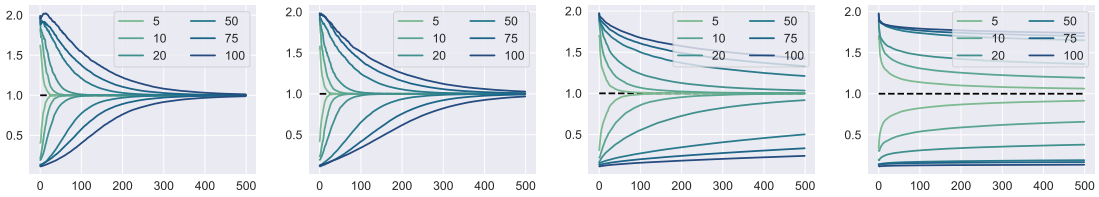
We implement the slice-matching scheme with source $\sigma = \mathcal{N}(0, \Sigma)$ and target $\mu = \mathcal{N}(0, \mathbf{I}_d)$ to illustrate our theoretical insights from Section 4. The updates are computed exactly following the explicit covariance recursion (18). We run the algorithm for different dimensions $d \in [5, 100]$ and step-size schedules $\gamma_k = (k+1)^{-\alpha}$ with $\alpha \in [0, 1)$. For each (d, α), we perform $N = 10$ independent runs (independent initializations of Σ), and we track the loss $SW_2^2(\sigma_k, \mu)$ (to verify convergence) and the extreme eigenvalues $\lambda_{\min}(\Sigma_k)$, $\lambda_{\max}(\Sigma_k)$.

Convergence and impact of (d, α). Figure 1 reports $SW_2^2(\sigma_k, \mu)$ as a function of the iteration k. For all tested dimensions d, the loss decreases, indicating convergence of the iterates toward the target measure. As d increases, the decay becomes slower, in agreement with our theoretical results, since the constants in our bounds scale polynomially with d. Similarly, the extreme eigenvalues converge to 1, which confirms that Σ_k becomes $\mathbf{I}_d$. Figure 1 also shows that smaller values of



(a) Setting (b) $\alpha = 0$ (c) $\alpha = 0.1$ (d) $\alpha = 0.51$ (e) $\alpha = 0.9$

Figure 1: Evolution of $SW_2^2(\sigma_k, \mu)$ when $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \mathbf{I}_d)$



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.51$ (d) $\alpha = 0.9$

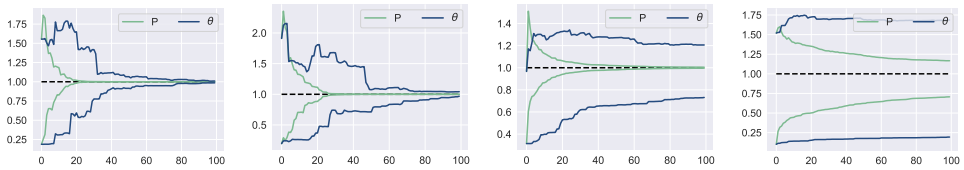
Figure 2: Minimum and maximum eigenvalues of Σ_k when $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \mathbf{I}_d)$

α (*i.e.*, more aggressive step sizes) yield faster empirical convergence, with $\alpha \in \{0, 0.1\}$ typically performing best. This behavior is not captured by our non-asymptotic analysis, derived for $\alpha > 0.5$. Extending the theory to values of α close to 0 remains an open problem.

Eigenvalue control. A key ingredient in our proof is to control $\lambda_{\min}(\Sigma_k)$ along the trajectory in order to verify Assumption A. In the isotropic target case $\mu = \mathcal{N}(0, \mathbf{I}_d)$, the theory predicts that once the second-moment bound $\mathrm{M}_2(\sigma_k) \leq \mathrm{M}_2(\mu)$ holds, which happens from the first iteration (Proposition 7), the eigenvalues remain uniformly bounded over k (Proposition 11). This behavior can be observed in Figure 2: the extreme eigenvalues settle in a fixed range from the first iteration. We emphasize that this behavior is due the moment-matching property inherent to the choice of an *orthonormal basis* P_{k+1} at each iteration. Another variant samples a single $\theta_{k+1} \in \mathbb{S}^{d-1}$ per iteration and updates only along that direction. The resulting extreme eigenvalues are shown in Figure 3, and exhibit larger fluctuations before stabilizing, which correlates with slower loss decay. The benefit of random orthonormal bases is consistent with recent work on sampling strategies in sliced OT (Sisouk et al., 2025).

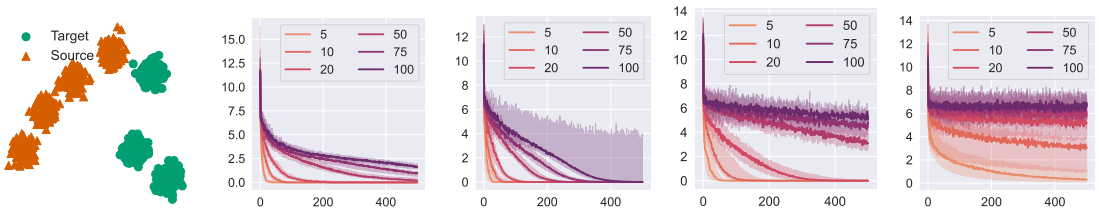
5.2. Beyond the Gaussian-to-Gaussian Setting

Figure 4 considers discrete empirical distributions of $n = 500$ samples. In each run, the source and target are sampled from a Gaussian mixture with randomly-generated mixture components. We plot $SW_2^2(\sigma_k, \mu)$ over iterations for $N = 10$ independent runs, across the same dimensions and step-size schedules as in the Gaussian setting. We observe the same trends: the loss decreases for all d, convergence slows down as d increases, and smaller values of α typically yield faster convergence. It is worth noting that $\alpha = 0.1$ outperforms $\alpha = 0$ in our experiments, which illustrates the interest of slice-matching algorithm (where (γ_k) is decaying) over IDT (where $\gamma_k = 1$). We provide additional experiments on empirical measures in Appendix F. While these discrete settings are not



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.51$ (d) $\alpha = 0.9$

Figure 3: Comparison of sampling strategies: single direction θ_{k+1} or orthonormal basis P_{k+1}. We report $\lambda_{\min}(\Sigma_k)$ and $\lambda_{\max}(\Sigma_k)$ with $\sigma = \mathcal{N}(0, \Sigma)$, $\mu = \mathcal{N}(0, \mathbf{I}_d)$, $d = 5$.



(a) Setting (b) $\alpha = 0$ (c) $\alpha = 0.1$ (d) $\alpha = 0.51$ (e) $\alpha = 0.9$

Figure 4: Evolution of $SW_2^2(\sigma_k, \mu)$ for discrete source and target distributions. The source and target samples are distributed from Gaussian mixtures.

covered by our theory, the empirical convergence suggests that regularity may hold more broadly, despite identified technical issues (Tanguy et al., 2025; Vauthier et al., 2025).

6. Conclusion and Perspectives

We established convergence rates for the slice-matching algorithm through Łojasiewicz-type inequalities for the Sliced-Wasserstein objective.We show that controlling the associated constants is tractable in the Gaussian (or elliptic) setting when sampling random orthonormal bases. A main limitation is that our explicit rate requires an isotropic Gaussian target, similarly to Cozzi and Santambrogio (2025). Extending the theory to general Gaussian targets and non-elliptic distributions remains open. A promising direction is to introduce regularization (*e.g.*, diffusive terms) to help maintain regularity along the dynamics (Liutkus et al., 2019; Tanguy et al., 2025; Chizat et al., 2025). Finally, our experiments show faster convergence with orthonormal bases of directions and step-size schedules $\gamma_k = 1/(k+1)^\alpha$ with small α. This behavior is not explained by our theorems and may require tools beyond decreasing-step stochastic approximation, for example Markov chains (Dieuleveut et al., 2020).

References

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. SIAM Journal on Mathematical Analysis, 43(2):904–924, 2011.

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. Journal of Machine Learning Research, 26(209):1–80, 2025.

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the bures-wasserstein manifold: dimension-free convergence of gradient descent. Advances in Neural Information Processing Systems, 34:22132–22145, 2021.

Luigi Ambrosio and Giuseppe Savaré. Gradient flows of probability measures. In Handbook of differential equations: evolutionary equations, volume 3, pages 1–136. Elsevier, 2007.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. Mathematics of operations research, 35(2):438–457, 2010.

Bernard Bercu and Jérémie Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. The Annals of Statistics, 49(2):968 – 987, 2021. doi: 10.1214/20-AOS1987.

Sergey Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances, volume 261. American Mathematical Society, 2019.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in wasserstein space. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

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

Benoît Bonnet. A pontryagin maximum principle in wasserstein spaces for constrained optimal control problems. ESAIM: Control, Optimisation and Calculus of Variations, 25:52, 2019.

Nicolas Bonnotte. Unidimensional and evolution methods for optimal transportation. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM review, 60(2):223–311, 2018.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417, 1991.

Luis A Caffarelli. The regularity of mappings with a convex potential. Journal of the American Mathematical Society, 5(1):99–104, 1992.

Luis A Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. Communications in Mathematical Physics, 214(3):547–563, 2000.

Guillaume Carlier, Alessio Figalli, Quentin Mérigot, and Yi Wang. Sharp comparisons between sliced and standard 1-Wasserstein distances, 2025.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In Conference on Learning Theory, 2020.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. arXiv:2407.18163, 3, 2024.

Lénaïc Chizat, Maria Colombo, and Xavier Fernández-Real. Convergence of drift-diffusion pdes arising as wasserstein gradient flows of convex functions. arXiv:2507.12385, 2025.

Kai Lai Chung. On a stochastic approximation method. The Annals of Mathematical Statistics, pages 463–483, 1954.

F. Coeurdoux, N. Dobigeon, and P. Chainais. Sliced-wasserstein normalizing flows: beyond maximum likelihood training. In Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, Oct. 2022.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 2016.

Giacomo Cozzi and Filippo Santambrogio. Long-time asymptotics of the sliced-wasserstein flow. SIAM Journal on Imaging Sciences, 18(1):1–19, 2025. doi: 10.1137/24M1656414.

Juan Antonio Cuesta and Carlos Matrán. Notes on the wasserstein metric in hilbert spaces. The Annals of Probability, pages 1264–1276, 1989.

Biwei Dai and Uros Seljak. Sliced iterative normalizing flows. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 2352–2364. PMLR, 18–24 Jul 2021.

I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. Schwing. Max-sliced wasserstein distance and its use for gans. In IEEE/CVF CVPR, 2019.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. The Annals of Statistics, 48(3):1348–1382, 2020.

Charles Dossal, Samuel Hurault, and Nicolas Papadakis. Optimization with first order algorithms. arXiv:2410.19506, 2024.

Chao Du, Tianbo Li, Tianyu Pang, Shuicheng Yan, and Min Lin. Nonparametric generative modeling with conditional sliced-Wasserstein flows. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 8565–8584. PMLR, 23–29 Jul 2023.

Marie Duflo. Algorithmes stochastiques, volume 23. Springer, 1996.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. arXiv:2301.11235, 2023.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. Mathematische Nachrichten, 147(1):185–203, 1990.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM journal on optimization, 23(4):2341–2368, 2013.

Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. ESAIM: Probability and Statistics, 23:841–873, 2019.

Louis Grenioux, Alain Oliviero Durmus, Eric Moulines, and Marylou Gabrié. On sampling with approximate transport maps. In Proceedings of the 40th International Conference on Machine Learning, 2023.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. The Annals of Statistics, 49(2):1166–1194, 2021.

Nicholas J Irons, Meyer Scetbon, Soumik Pal, and Zaid Harchaoui. Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. In International Conference on Artificial Intelligence and Statistics, pages 10161–10195. PMLR, 2022.

Li Jiang, Xiao Li, Andre Milzarek, and Junwen Qiu. A Generalized Version of Chung's Lemma and its Applications, 2024.

Parnian Kassraie, Aram-Alexandre Pooladian, Michal Klein, James Thornton, Jonathan Niles-Weed, and Marco Cuturi. Progressive entropic optimal transport solvers. Advances in Neural Information Processing Systems, 37:19561–19590, 2024.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. Survey Certification.

Jun Kitagawa and Asuka Takatsu. Sliced optimal transport: is it a suitable replacement?, 2024.

Benoît Kloeckner. A geometric study of wasserstein spaces: Euclidean spaces. Annali della Scuola Normale Superiore di Pisa-Classe di Scienze, 9(2):297–323, 2010.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In International Conference on Learning Representations, 2018.

Krzysztof Kurdyka, Tadeusz Mostowski, and Adam Parusiński. Proof of the gradient conjecture of r. thom. Annals of Mathematics, pages 763–792, 2000.

Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-order conditions for optimization in the wasserstein space. SIAM Journal on Mathematics of Data Science, 7(1):274–300, 2025.

Shiying Li and Caroline Moosmüller. Approximation properties of slice-matching operators. Sampling Theory, Signal Processing, and Data Analysis, 22(1):15, 2024.

Shiying Li, Caroline Moosmueller, and Yongzhe Wang. Measure transfer via stochastic slicing and matching. arXiv:2307.05705, 2023.

Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. In The Thirteenth International Conference on Learning Representations, 2025.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In International Conference on machine learning, 2019.

Jianyu Ma. Absolute continuity of wasserstein barycenters on manifolds with a lower ricci curvature bound. arXiv:2310.13832, 2023.

Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast Optimal Transport through Sliced Generalized Wasserstein Geodesics. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

Tudor Manole, Sivaraman Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the Sliced Wasserstein distance. Electronic Journal of Statistics, 16(1):2252 – 2345, 2022. doi: 10.1214/22-EJS2001.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via Measure Transport: An Introduction. Springer International Publishing, 2016.

Robert J McCann. A convexity principle for interacting gases. Advances in mathematics, 128(1): 153–179, 1997.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Advances in neural information processing systems, 24, 2011.

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, 32, 2019.

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, 33:20802–20812, 2020.

Alexander M Ostrowski. A quantitative formulation of sylvester's law of inertia. Proceedings of the National Academy of Sciences, 45(5):740–744, 1959.

Sangmin Park and Dejan Slepčev. Geometry and analytic properties of the sliced wasserstein space. Journal of Functional Analysis, 289(7):110975, 2025.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

François Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. Computer Vision and Image Understanding, 107(1-2):123–137, 2007.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In International conference on scale space and variational methods in computer vision, pages 435–446. Springer, 2011.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein, editors, Scale Space and Variational Methods in Computer Vision, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Optimizing methods in statistics, pages 233–257. Elsevier, 1971.

R. Tyrrell Rockafellar. Convex analysis. Princeton University Press, 1970.

Filippo Santambrogio. Optimal transport for applied mathematicians, volume 87. Springer, 2015.

Keanu Sisouk, Julie Delon, and Julien Tierny. A User's Guide to Sampling Strategies for Sliced Optimal Transport. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. Survey Certification.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

Eloi Tanguy. Convergence of SGD for training neural networks with sliced wasserstein losses. Transactions on Machine Learning Research, 2023. ISSN 2835-8856.

Eloi Tanguy, Rémi Flamary, and Julie Delon. Reconstructing discrete measures from projections. consequences on the empirical sliced wasserstein distance. Comptes Rendus. Mathématique, 362 (G10):1121–1129, 2024.

Eloi Tanguy, Rémi Flamary, and Julie Delon. Properties of discrete sliced Wasserstein losses. Mathematics of Computation, 94(353):1411–1465, 2025.

Christophe Vauthier, Quentin Mérigot, and Anna Korba. Properties of Wasserstein Gradient Flows for the Sliced-Wasserstein Distance. arXiv:2502.06525, 2025.

Cédric Villani. Optimal transport: old and new, volume 338. Springer, 2008.

Sven Wang and Youssef Marzouk. On minimax density estimation via measure transport. arXiv:2207.10231, 2022.

Douglas P Wiens. On moments of quadratic forms in non-spherically distributed variables. Statistics, 23(3):265–270, 1992.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. arXiv:1803.06573, 2018.

Appendix A. Reminders on Wasserstein space

This appendix gathers existing results useful for optimization over the space of probability distributions. For further details, we refer the interesting reader to the classical references Ambrosio and Savaré (2007); Santambrogio (2015). First, recall that, for $\psi^c(y) = \inf_x \{\frac{1}{2}\|x-y\|^2 - \psi(x)\}$ the c-transform of ψ, the dual of Kantorovich OT problem writes

$$W_2^2(\alpha,\beta) = \sup_{\psi \in L^1(\alpha)} \int \psi \mathrm{d}\alpha + \int \psi^c \mathrm{d}\beta. \tag{22}$$

The solution of the latter is called the Kantorovich potential, and it is unique (up to translations) under finiteness of second-order moments, with α giving no mass to $d-1$ surfaces (Santambrogio, 2015, Theorem 1.22).

A.1. Curves and convexity in Wasserstein space

The Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is the space of square-integrable probability distributions endowed with the Wasserstein distance W_2. A first way to construct an absolutely continuous curve between two measures σ_0 and σ_1 is the flat interpolation, given, for $t \in [0,1]$, by

$$\sigma_t = (1-t)\sigma_0 + t\sigma_1. \tag{23}$$

This convex combination between densities ignores the geometry induced by the Wasserstein distance. In contrast, denoting by $T_{\sigma_0}^{\sigma_1}$ the OT map from σ_0 to σ_1, another interpolation is given by

$$\sigma_t = \big((1-t)\mathrm{Id} + tT_{\sigma_0}^{\sigma_1}\big)_\sharp \sigma_0. \tag{24}$$

Due to the fact that $(1-t)\mathrm{Id} + tT_{\sigma_0}^{\sigma_1}$ is the gradient of a convex function, it is the solution of Monge OT problem (Brenier, 1991; Cuesta and Matrán, 1989). Hence, σ_t corresponds to the shortest path between σ_0 and σ_1, in the sense that

$$\forall\, 0 \le s \le t \le 1, \qquad W_2(\sigma_s,\sigma_t) = (t-s)W_2(\sigma_0,\sigma_1).$$

While (23) corresponds to a mixture model between σ_0 and σ_1, the interpolant (24) is more of a barycenter (Agueh and Carlier, 2011; Rabin et al., 2011) and it is a building block for gradient flows in the Wasserstein space (Ambrosio and Savaré, 2007). Interestingly enough, $W_2^2(\cdot,\sigma)$ is strictly convex along (23) as soon as σ is absolutely continuous (Santambrogio, 2015, Proposition 7.19). It is not hard to see that the same property holds for the Sliced-Wasserstein distance, with arguments reminiscent to the ones of Ma (2023, Proposition 2.10) for Wasserstein barycenters. Such convexity along (23) must be understood with respect to the 2-norm between densities. The analog along (24), with respect to the Wasserstein distance, writes as follows.

Definition 6 *$\mathcal{F}$ is geodesically α-convex if, for all $\sigma_0,\sigma_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\sigma_t = ((1-t)Id + tT_{\sigma_0}^{\sigma_1})_\sharp\sigma_0$,*

$$\mathcal{F}(\sigma_t) \le (1-t)\mathcal{F}(\sigma_0) + t\mathcal{F}(\sigma_1) - \frac{\alpha}{2}t(1-t)W_2^2(\sigma_0,\sigma_1).$$

Unfortunately, the reverse inequality holds for $W_2^2(\cdot,\sigma)$ in general dimension (Ambrosio and Savaré, 2007, Theorem 7.3.2), and a fortiori for the Sliced-Wasserstein distance up to integration

over the projection directions (Vauthier et al., 2025, Appendix A.5). These facts are discussed in Lemma 3.

The situation is very different in dimension $d = 1$, due to the particular properties of $(\mathcal{P}_2(\mathbb{R}), W_2)$. In this setting, the composition of OT maps preserves their monotonicity (hence the optimality) and W_2 rewrites with Q_0, Q_1 the quantile functions of σ_0, σ_1:

$$W_2^2(\sigma_0, \sigma_1) = \int_0^1 \|Q_0(t) - Q_1(t)\|^2 dt \tag{25}$$

or, equivalently, $W_2^2(\sigma_0, \sigma_1) = \|T_\rho^{\sigma_0} - T_\rho^{\sigma_1}\|_\rho^2$ for any pivot measure $\rho \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. As a byproduct, the geodesics in (24) coincide with the generalized geodesics $\sigma_t = \big((1-t)T_\rho^{\sigma_0} + tT_\rho^{\sigma_1}\big)_\sharp \rho$, and one can find in Ambrosio and Savaré (2007, Chapter 9) that they verify the generalized parallelogram rule

$$W_2^2(\sigma_t, \sigma) = (1-t)W_2^2(\sigma_0, \sigma) + tW_2^2(\sigma_1, \sigma) - t(1-t)W_2^2(\sigma_0, \sigma_1). \tag{26}$$

This can be easily verified by expanding the square in $W_2^2(\sigma_t, \sigma)$ via (25) and using the tricks $t^2 = t - t(1-t)$ and $(1-t)^2 = (1-t) - t(1-t)$, as in Kloeckner (2010, Proposition 4.1). Next, we turn to differentiation along geodesics. Unfortunately, (26) does not imply the same parallelogram identity for the Sliced-Wasserstein distance, as it would require to identify a path σ_t in $\mathbb{R}^d$ along the map $\overline{T} : x \mapsto \int_\theta T_{\sigma^\theta}^{\mu^\theta}(x) d\mathcal{U}(\theta)$ with all projected generalized geodesics σ_t^θ, which is not true.

A.2. Differentiation along geodesics

We borrow the differential structure of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ as described in *e.g.*, Bonnet (2019); Bonet et al. (2024); Lanzetti et al. (2025). The tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at σ is defined by

$$\mathcal{T}_\sigma = \overline{\{\nabla\psi : \psi \in C_c^\infty(\mathbb{R}^d)\}},$$

where the closure is taken with respect to $\mathbf{L}^2(\sigma)$, the set of σ-square integrable functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, and where $C_c^\infty(\mathbb{R}^d)$ is the set of infinitely differentiable functions with compact support. Consider $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. At any $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{F}(\sigma) < +\infty$, the Wasserstein gradient $\nabla_{W_2}\mathcal{F}(\sigma)$ is the unique vector in $\mathcal{T}_\sigma$ verifying, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and any optimal coupling $\gamma \in \Pi(\sigma, \mu)$[2],

$$\mathcal{F}(\mu) = \mathcal{F}(\sigma) + \int \langle \nabla_{W_2}\mathcal{F}(\sigma)(x), y - x\rangle \mathrm{d}\gamma(x, y) + o(W_2(\sigma, \mu)). \tag{27}$$

One way to compute the Wasserstein gradient is by taking $\nabla_{W_2}\mathcal{F}(\sigma) = \nabla\frac{\delta\mathcal{F}}{\delta\sigma}(\sigma)$. for $\frac{\delta\mathcal{F}}{\delta\sigma}(\sigma)$ the first variation (Santambrogio, 2015, Definition 7.12) defined as follows. Firstly, a measure $\rho \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ is *regular* for $\mathcal{F}$ if, for every $\overline{\rho} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ with L^∞ density and compact support, $\mathcal{F}((1-t)\rho + t\overline{\rho}) < +\infty$ for every $t \in [0, 1]$. With this at hand, if ρ is regular for $\mathcal{F}$, the first variation $\frac{\delta\mathcal{F}}{\delta\rho}(\rho)$ verifies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho + t\xi)_{|t=0} = \lim_{t\to 0} \frac{\mathcal{F}(\rho + t\xi) - \mathcal{F}(\rho)}{t} = \int \frac{\delta\mathcal{F}}{\delta\rho}(\rho)\mathrm{d}\xi,$$

for all $\xi = \overline{\rho} - \rho$ with $\overline{\rho} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ with L^∞ density and compact support. The following useful remark is taken from (Santambrogio, 2015, Remark 7.14).

2. The set of couplings between σ and μ is $\Pi(\sigma, \mu) = \{\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : \pi(A \times \mathbb{R}^d) = \sigma(A)\,,\, \pi(\mathbb{R}^d \times B) = \mu(B)\}$.

Remark 7 *For $\sigma \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $\mathcal{F} : \rho \mapsto W_2^2(\rho, \sigma)$, any $\rho \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ is regular if and only if $\mathcal{F}(\rho) < +\infty$. Indeed, for every $\overline{\rho} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, $\mathcal{F}((1-t)\rho + t\overline{\rho}) \leq (1-t)\mathcal{F}(\rho) + t\mathcal{F}(\overline{\rho})$ by strict convexity (Santambrogio, 2015, Proposition 7.19). Hence, as soon as $\overline{\rho}$ is compactly supported, $\mathcal{F}(\overline{\rho}) < +\infty$ and ρ is regular if $\mathcal{F}(\rho) < +\infty$. The reciprocal is immediate by taking $t = 0$ in the definition of a regular measure.*

When considering the Wasserstein distance $\sigma \mapsto W_2^2(\sigma, \mu)$, the first variation is ψ, the Kantorovich potential that is solution of (22) (Proposition 7.17, Santambrogio, 2015), and a similar statement holds for $\sigma \mapsto SW_2^2(\sigma, \mu)$ (Cozzi and Santambrogio, 2025). We discuss this in Proposition 6.

Appendix B. Proofs of Sections 2 and 3

In this section, we detail the properties of the functional to be minimized. We discuss differentiability and critical points, before turning to smoothness and boundedness of the gradient, the latter being a byproduct of boundedness of moments along the iterations.

B.1. Differentiability, critical points

The next proposition describes Wasserstein gradients of our sliced objective, as previously provided in Bonnotte (2013); Cozzi and Santambrogio (2025). We also detail simple properties of the gradient norm, that are important with the purpose of SGD.

Proposition 6 *Given that $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ is compactly supported, the Wasserstein gradients of $\mathscr{F}(\cdot, \theta)$ and $\mathscr{F}$ at any $\sigma \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ compactly supported are given by*

$$\nabla_{W_2}\mathscr{F}(\sigma, P) = Id - T_{\sigma,P} \qquad \text{and} \qquad \nabla_{W_2}\mathscr{F}(\sigma) = d\int \theta(Id - T_{\sigma^\theta}^{\mu^\theta}) \circ \pi_\theta \, \mathrm{d}\mathcal{U}(\theta),$$

where $T_{\sigma^\theta}^{\mu^\theta}$ denotes the one-dimensional OT map pushing σ^θ to μ^θ. It follows that

$$\|\nabla_{W_2}\mathscr{F}(\sigma, P)\|_\sigma^2 = \sum_{\ell=1}^d W_2^2(\sigma^{\theta_\ell}, \mu^{\theta_\ell}) \qquad \text{and} \qquad \|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 \leq dSW_2^2(\sigma, \mu),$$

that is

$$\|\nabla_{W_2}\mathscr{F}(\sigma, P)\|_\sigma^2 = 2\mathscr{F}(\sigma, P) \qquad \text{and} \qquad \|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 \leq 2\mathscr{F}(\sigma).$$

In fact, this can be refined in the following decomposition, for $\overline{T}_\sigma = \mathbb{E}_P[T_{\sigma,P}]$, [3]

$$2\mathscr{F}(\sigma) = \|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2 + \mathbb{E}_P[\|\overline{T}_\sigma - T_{\sigma,P}\|_\sigma^2].$$

Proof For a given basis P, the Wasserstein gradient of $\sigma \mapsto \mathscr{F}(\sigma, P)$ is given by the euclidean gradient of its first variation, i.e., $\nabla_{W_2}\mathscr{F}(\sigma, P) = \nabla\frac{\delta\mathscr{F}}{\delta\sigma}(\sigma, P)$, as recalled in Appendix A.2. From Santambrogio (Proposition 7.17, 2015), the first variation of $\sigma \mapsto W_2^2(\sigma^\theta, \mu^\theta)$ is given by $\varphi_\theta(\langle \cdot, \theta \rangle)$, for φ_θ the first Kantorovich potential φ_θ for the OT problem from σ^θ to μ^θ (assuming compactness of the underlying supports). Thus, the first variation of $\mathscr{F}(\sigma, P) = \sum_{\ell=1}^d W_2^2(\sigma^{\theta_\ell}, \mu^{\theta_\ell})$

3. Equivalently, $dSW_2^2(\sigma, \mu) = \|\mathrm{Id} - \overline{T}_\sigma\|_\sigma^2 + \mathbb{E}_P\left[\|\overline{T}_\sigma - T_{\sigma,P}\|_\sigma^2\right]$.

is $\sum_{\ell=1}^{d} \varphi_{\theta_\ell}(\langle \cdot, \theta_\ell \rangle)$, and the euclidean gradient is given through $\nabla \varphi_{\theta_\ell}(\langle x, \theta_\ell \rangle) = \theta_\ell (x^\top \theta_\ell - T_{\sigma^{\theta_\ell}}^{\mu^{\theta_\ell}}(x^\top \theta_\ell))$. The first result directly follows as

$$\nabla_{W_2} \mathscr{F}(\sigma, P) = \sum_{\ell=1}^{d} \theta_\ell (\mathrm{Id} - T_{\sigma^{\theta_\ell}}^{\mu^{\theta_\ell}}) \circ \pi_{\theta_\ell} = \mathrm{Id} - T_{\sigma,P}.$$

Regarding the integrated version over the directions, the first variation of $\mathscr{F}(\sigma) = dSW_2^2(\sigma, \mu)$ is

$$x \mapsto d \int \varphi_\theta(x^\top \theta) \mathrm{d}\mathcal{U}(\theta), \tag{28}$$

as stated in Cozzi and Santambrogio (2025). The detail of this calculus requires interchanging a limit and an integral, because, by definition,

$$\int \frac{\delta \mathscr{F}}{\delta \sigma}(\sigma) \mathrm{d}\xi = \lim_{t \to 0} \frac{\mathscr{F}(\sigma + t\xi) - \mathscr{F}(\sigma)}{t} = \lim_{t \to 0} \mathbb{E}_P \frac{\mathscr{F}(\sigma + t\xi, P) - \mathscr{F}(\sigma, P)}{t},$$

for all $\xi = \overline{\rho} - \rho$ with $\overline{\rho} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ with L^∞ density and compact support. Under compact assumptions, this can be treated as in the last step of the proof of Santambrogio (2015, Proposition 7.17). A direct consequence of (28) is that $\nabla_{W_2} \mathscr{F}(\sigma) = d \int \theta (\mathrm{Id} - T_{\sigma^\theta}^{\mu^\theta}) \circ \pi_\theta \, d\mathcal{U}(\theta)$. Now,

$$\|\nabla_{W_2} \mathscr{F}(\sigma, P)\|_\sigma^2 = \sum_{\ell=1}^{d} \|\langle \cdot, \theta_\ell \rangle - T_{\sigma^{\theta_\ell}}^{\mu^{\theta_\ell}}(\langle \cdot, \theta_\ell \rangle)\|_\sigma^2 = \sum_{\ell=1}^{d} W_2^2(\sigma^{\theta_\ell}, \mu^{\theta_\ell}) = 2\mathscr{F}(\sigma, P),$$

so that

$$\mathbb{E}_P\big(\|\nabla_{W_2} \mathscr{F}(\sigma, P)\|_\sigma^2\big) = \mathbb{E}_P\big(2\mathscr{F}(\sigma, P)\big) = 2\mathscr{F}(\sigma). \tag{29}$$

Also, Jensen's inequality implies

$$\|\nabla_{W_2} \mathscr{F}(\sigma)\|_\sigma^2 = \|\mathbb{E}_P \nabla_{W_2} \mathscr{F}(\sigma, P)\|_\sigma^2 \leq \mathbb{E}_P \|\nabla_{W_2} \mathscr{F}(\sigma, P)\|_\sigma^2 \leq 2\mathscr{F}(\sigma).$$

Finally, recall the decomposition

$$\nabla_{W_2} \mathscr{F}(\sigma, P) = \nabla_{W_2} \mathscr{F}(\sigma) + \overline{T}_\sigma - T_{\sigma,P}.$$

Taking the square norm, developing the square and using that $\mathbb{E}_P[\overline{T}_\sigma - T_{\sigma,P}] = 0$, one obtains that

$$\mathbb{E}_P\big(\|\nabla_{W_2} \mathscr{F}(\sigma, P)\|_\sigma^2\big) = \|\nabla_{W_2} \mathscr{F}(\sigma)\|_\sigma^2 + \mathbb{E}_P \|\overline{T}_\sigma - T_{\sigma,P}\|_\sigma^2.$$

Combining the above with (29), provides the desired decomposition. ■

Remark 8 *Without compacity, Proposition 6 does not hold, but one can still define $\nabla_{W_2} \mathscr{F}(\sigma)$ directly through $\nabla_{W_2} \mathscr{F}(\sigma) = \int \theta (Id - T_{\sigma^\theta}^{\mu^\theta}) \circ \pi_\theta \, d\mathcal{U}(\theta)$. In this case, $\nabla_{W_2} \mathscr{F}(\sigma)$ belongs to the subdifferential of $\mathscr{F}(\sigma)$ (Proposition 4.7(b), Vauthier et al., 2025).*

Vauthier et al. (2025) describe different possible notions of critical points, including the following.

Definition 9 (Definition 4.2 from Vauthier et al. (2025)) *A measure σ is a barycentric Lagrangian critical point for $SW_2^2(\cdot, \mu)$ if,*

$$\frac{1}{d}x = \int_{\mathbb{S}^{d-1}} \theta T_{\sigma^\theta}^{\mu^\theta}(x^\top\theta)\mathrm{d}\mathcal{U}(\theta) \qquad \text{for } \sigma\text{-a.e. } x\,,$$

where $T_{\sigma^\theta}^{\mu^\theta}$ corresponds to the OT map from σ^θ to μ^θ, the pushforward measures of σ and μ by $x \mapsto \theta^\top x$.

With our notations, a critical point of $\sigma \mapsto \mathscr{F}(\sigma)$ verifies $\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma = 0$, hence

$$\int \|\frac{x}{d} - \int \theta T_{\sigma^\theta}(x^\top\theta)d\mathcal{U}(\theta)\|^2 \mathrm{d}\sigma(x) = 0,$$

and it is a barycentric Lagrangian critical point for $SW_2^2(\cdot, \mu)$[4]. We stress that this only implies $T_{\sigma^\theta}^{\mu^\theta} = \mathrm{Id}$ and $\sigma^\theta = \mu^\theta$ on average $w.r.t.$ θ, which is weaker than $SW_2(\sigma, \mu) = 0$ where $T_{\sigma^\theta}^{\mu^\theta} = \mathrm{Id}$ for $\mathcal{U}$-$a.e.$ $\theta \in \mathbb{S}^{d-1}$. Although critical points of $\mathscr{F}$ may differ from μ (Vauthier et al., 2025), the next lemma describes conditions under which it must coincide.

Lemma 2 (Lemma 5.7.2 from Bonnotte (2013)) *Suppose that the target measure $\mu \in \mathcal{P}_{2,ac}(B(0, r))$ has a strictly positive density. Then, $\sigma = \mu$ if and only if $\nabla_{W_2}\mathscr{F}(\sigma) = 0$.*

Lemma 2 provides assumptions under which convergence towards a critical point implies convergence towards the target measure μ.

B.2. Smoothness

Lemma 3 *For $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\mathcal{F}(\sigma) = \frac{1}{2}SW_2^2(\sigma, \mu) = \frac{1}{d}\mathscr{F}(\sigma)$. Let $T_0, T_1 \in \mathbf{L}^2(\sigma)$ such that $\sigma_0 = T_0\sharp\sigma$ and $\sigma_1 = T_1\sharp\sigma$. For $t \in (0, 1)$ and $\sigma_t = ((1-t)T_0 + tT_1)_\sharp\sigma$,*

$$SW_2^2(\sigma_t, \mu) \geq (1-t)SW_2^2(\sigma_0, \mu) + tSW_2^2(\sigma_1, \mu) - t(1-t)\frac{1}{d}\|T_0 - T_1\|_\sigma^2, \tag{30}$$

and

$$\langle \nabla_{W_2}\mathcal{F}(\sigma_0)\circ T_0 - \nabla_{W_2}\mathcal{F}(\sigma_1)\circ T_1, T_0 - T_1\rangle_\sigma \leq \frac{1}{d}\|T_0 - T_1\|_\sigma^2. \tag{31}$$

Besides, for any $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, for any $T \in \mathbf{L}^2(\sigma)$,

$$SW_2^2(T_\sharp\sigma, \mu) \leq SW_2^2(\sigma, \mu) + 2\langle \nabla_{W_2}\mathcal{F}(\sigma), T - Id\rangle_\sigma + \frac{1}{d}\|T - Id\|_\sigma^2. \tag{32}$$

Lemma 3 is simply a rewriting of results from Vauthier et al. (2025, Proposition 4.7), with (31) being a well-known equivalent characterization for smoothness (Zhou, 2018).

Proof Define $\mathcal{F}_\sigma : T \mapsto \mathcal{F}(T_\sharp\sigma)$ on $(\mathbf{L}^2(\sigma), \|\cdot\|_\sigma)$. One has that $\nabla\mathcal{F}_\sigma(T) = \nabla_{W_2}\mathcal{F}(T_\sharp\sigma)\circ T$ (Bonet et al., 2024, Proposition 1), so that smoothness results on $\mathcal{F}_\sigma$ are equivalent to that on $\mathscr{F}$, except that the linear structure of $(\mathbf{L}^2(\sigma), \|\cdot\|_\sigma)$ is easier to deal with than $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

4. One might note that $\int \theta\theta^\top x\mathrm{d}\mathcal{U}(\theta) = \int \theta\theta^\top \mathrm{d}\mathcal{U}(\theta)x = x/d$.

The inequality (31) is a rewriting of Vauthier et al. (2025, Proposition 4.7), that itself follows from the semi-concavity of Wasserstein distances along generalized geodesics (Ambrosio and Savaré, 2007, Theorem 7.3.2). A change-of-variables ($T - \mathrm{Id} = \xi$) in (Vauthier et al., 2025, Proposition 4.7(a)) gives us that

$$G_\sigma : T \mapsto \frac{1}{2d}\|T - \mathrm{Id}\|_\sigma^2 - \mathcal{F}_\sigma(T)$$

is convex on $(\mathbf{L}^2(\sigma), \|\cdot\|_\sigma)$. But note that $\nabla G_\sigma(T) = \frac{1}{d}(T - \mathrm{Id}) - \nabla\mathcal{F}_\sigma(T)$. Hence, first-order conditions for convexity (Bonet et al., 2024, Proposition 13) applied to F_σ yield

$$\langle \nabla G_\sigma(T_0) - \nabla G_\sigma(T_1), T_0 - T_1\rangle_\sigma \geq 0,$$

which directly implies (31). Besides, one can find in Vauthier et al. (2025, Appendix B.6), namely the equations (140) and (155), that, for any $\sigma, \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

(a) for $\xi_0, \xi_1 \in \mathbf{L}^2(\sigma)$, $\xi_t = (1-t)(\mathrm{Id} + \xi_0) + t(\mathrm{Id} + \xi_1)$ and $\sigma_t = (\xi_t)_\sharp\sigma$:

$$SW_2^2(\sigma_t, \mu) \geq (1-t)SW_2^2(\sigma_0, \mu) + tSW_2^2(\sigma_1, \mu) - t(1-t)\frac{1}{d}\|\xi_0 - \xi_1\|_\sigma^2,$$

(b) for $\xi \in \mathbf{L}^2(\sigma)$: $SW_2^2((\mathrm{Id} + \xi)_\sharp\sigma, \mu) \leq SW_2^2(\sigma, \mu) + 2\langle \nabla_{W_2}\mathcal{F}(\sigma), \xi\rangle_\sigma + \frac{1}{d}\|\xi\|_\sigma^2$.

When replacing $T_i = \mathrm{Id} + \xi_i$ and $T = \mathrm{Id} + \xi$, one recovers immediately (30) and (32). ■

Corollary 10 *$\mathscr{F}$ is 1-smooth with respect to the Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^d)$, i.e., for any $\sigma_1, \sigma_2 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathscr{F}(\sigma_2) \leq \mathscr{F}(\sigma_1) + \langle \nabla_{W_2}\mathscr{F}(\sigma_1), T_{\sigma_1}^{\sigma_2} - Id\rangle_{\sigma_1} + \frac{1}{2}W_2^2(\sigma_1, \sigma_2)\,.$$

Proof For any $\sigma_1, \sigma_2 \in \mathcal{P}_2(\mathbb{R}^d)$, if the OT map $T_{\sigma_1}^{\sigma_2}$ from σ_1 to σ_2 exists, then $\|T_{\sigma_1}^{\sigma_2} - \mathrm{Id}\|_{\sigma_1}^2 = W_2^2(\sigma_1, \sigma_2)$. The final result follows from (32). ■

The following lemma is new, although it is not required for our main results. It resembles a well-known smoothness property, but we stress that, even in Euclidean settings, it is not equivalent to the previous inequalities (Zhou, 2018).

Lemma 4 *Fix $\sigma_1, \sigma_2 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. If the density function of the target μ is strictly larger than $1/\kappa > 0$ on its compact domain, then*

$$\|\nabla_{W_2}\mathscr{F}(\sigma_1) - \nabla_{W_2}\mathscr{F}(\sigma_2)\|_\lambda^2 \leq 2\kappa\, SW_1(\sigma_1, \sigma_2),$$

for λ the Lebesgue measure.

Proof By Jensen's inequality,

$$\begin{aligned}\|\nabla_{W_2}\mathscr{F}(\sigma_1) - \nabla_{W_2}\mathscr{F}(\sigma_2)\|_\lambda^2 &\leq \int\int |T_{\sigma_1^\theta}^{\mu^\theta}(x^\top\theta) - T_{\sigma_2^\theta}^{\mu^\theta}(x^\top\theta)|^2 \mathrm{d}\lambda(x)\mathrm{d}\mathcal{U}(\theta),\\ &\leq \int\int |C_{\mu^\theta}^{-1} \circ C_{\sigma_1^\theta}(x^\top\theta) - C_{\mu^\theta}^{-1} \circ C_{\sigma_2^\theta}(x^\top\theta)|^2 \mathrm{d}\lambda(x)\mathrm{d}\mathcal{U}(\theta),\end{aligned}$$

for C_ρ the univariate distribution function of $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. For all $\theta \in \mathbb{S}^{d-1}$, the quantile function C_{μ^θ} is κ-lipschitz with $1/\kappa$ the essential infimum of the density of μ on its domain, (Bobkov and Ledoux, 2019). Then, the result follows from

$$\|\nabla_{W_2}\mathscr{F}(\sigma_1) - \nabla_{W_2}\mathscr{F}(\sigma_2)\|_\lambda^2 \leq 2\kappa \int\int |C_{\sigma_1^\theta}(x^\top\theta) - C_{\sigma_2^\theta}(x^\top\theta)|\mathrm{d}\lambda(x)\mathrm{d}\mathcal{U}(\theta).$$

■

B.3. Moments are bounded

An important assumption when dealing with stochastic algorithms is the boundedness of the gradient norm. Here, a direct consequence of Proposition 6 is that, using respectively the definition of the Haar measure and Jensen's inequality,

$$\mathbb{E}_P\|\nabla_{W_2}\mathscr{F}(\sigma, P)\|_\sigma^2 = \mathbb{E}_P\sum_{\ell=1}^d W_2^2(\sigma^{\theta_\ell}, \mu^{\theta_\ell}) = 2\mathscr{F}(\sigma) \quad \text{and} \quad \|\nabla_{W_2}\mathscr{F}(\sigma)\|^2 \leq 2\mathscr{F}(\sigma).$$

Hence, the gradient norm is bounded as long as the objective $\mathscr{F}(\sigma)$ is. We now show that the second-order moments remain bounded along the IDT iterations, a fact that implies a bound on $(\mathscr{F}(\sigma_k))_k$. Bounds on the moments along the Sliced-Wasserstein flow were proved in Cozzi and Santambrogio (2025), in a continuous-time setting, whereas we deal with discrete step sizes (γ_k). Denote by $\mathrm{M}_2(\rho) = \int \|\cdot\|^2 d\rho$ the second-order moment of a probability distribution $\rho \in \mathcal{P}_2(\mathbb{R}^d)$.

Proposition 7 *Moments are bounded by $\mathrm{M}_2(\mu)$ along the IDT iterations (4). In other words, for any $k \geq 1$, we have*

$$\mathrm{M}_2(\sigma_k) \leq \mathrm{M}_2(\mu).$$

Consequently, for $M = 4\mathrm{M}_2(\mu)$,

$$W_2^2(\sigma_k, \mu) \leq M \qquad \textit{and} \qquad SW_2^2(\sigma_k, \mu) \leq \frac{M}{d}.$$

Proof This result is a consequence of the moment-matching property of sliced maps. Indeed, as shown in Li and Moosmüller (2024, Proposition 3.6), for all $k \geq 0$,

$$\begin{aligned}\int \|T_{\sigma_k, P_{k+1}}(x)\|^2\mathrm{d}\sigma_k(x) &= \int \|\sum_{\ell=1}^d \theta_\ell t_{\theta_\ell}(\theta_\ell^\top x)\|^2\mathrm{d}\sigma_k(x) = \sum_{\ell=1}^d \int \|t_{\theta_\ell}(\theta_\ell^\top x)\|^2\mathrm{d}\sigma_k(x),\\ &= \sum_{\ell=1}^d \mathrm{M}_2(\mu^{\theta_\ell}) = \sum_{\ell=1}^d \int \langle y, \theta_\ell\rangle^2\mathrm{d}\mu(y) = \int \|y\|^2\mathrm{d}\mu(y) = \mathrm{M}_2(\mu).\end{aligned} \tag{33}$$

With this at hand, we proceed by induction. At initialization, for $k = 1$, we have that $\gamma_1 = 1$ and $\mathrm{M}_2(\sigma_1) = \int \|T_{\sigma_0, P_1}(x)\|^2 d\sigma_0(x) = \mathrm{M}_2(\mu)$. For the induction step, assume that there exists an index $k \in \mathbb{N}^*$ such that $\mathrm{M}_2(\sigma_k) \leq \mathrm{M}_2(\mu)$. Then, by convexity of $\|\cdot\|^2$,

$$\begin{aligned}\mathrm{M}_2(\sigma_{k+1}) &= \int \|x\|^2\mathrm{d}\sigma_{k+1}(x) = \int \|(1-\gamma_k)x + \gamma_k T_{\sigma_k, P_{k+1}}(x)\|^2\mathrm{d}\sigma_k(x),\\ &\leq (1-\gamma_k)\mathrm{M}_2(\sigma_k) + \gamma_k \int \|T_{\sigma_k, P_{k+1}}(x)\|^2\mathrm{d}\sigma_k(x).\end{aligned} \tag{34}$$

Plugging (33) in (34) and invoking the induction hypothesis that $\mathrm{M}_2(\sigma_k) \leq \mathrm{M}_2(\mu)$, the desired result on the moment boundedness follows:

$$\mathrm{M}_2(\sigma_{k+1}) \leq (1-\gamma_k)\mathrm{M}_2(\sigma_k) + \gamma_k \mathrm{M}_2(\mu) \leq (1-\gamma_k)\mathrm{M}_2(\mu) + \gamma_k \mathrm{M}_2(\mu) \leq \mathrm{M}_2(\mu). \tag{35}$$

Next, to obtain the bound on the Wasserstein distance $W_2^2(\sigma_k, \mu)$, let us call T^* the OT map from σ_k to μ. Young's inequality for products together with the change-of-variable $\mathbb{E}_{X\sim\sigma_k}(\|T^*(X)\|^2) = \mathbb{E}_{Y\sim\mu}(\|Y\|^2)$ leads to

$$\begin{aligned} W_2^2(\sigma_k, \mu) &= \mathbb{E}_{X\sim\sigma_k}[\|X - T^*(X)\|^2] \\ &= \mathbb{E}[\|X\|^2] + \mathbb{E}[\|T^*(X)\|^2] - 2\mathbb{E}[\langle X, T^*(X)\rangle] \\ &\leq 2\left(\mathbb{E}[\|X\|^2] + \mathbb{E}[\|T^*(X)\|^2]\right) \\ &\leq 2(\mathrm{M}_2(\sigma_k) + \mathrm{M}_2(\mu))\,, \end{aligned}$$

that is the desired result. ■

B.4. Standard proofs for non-convex smooth optimization

The next two demonstrations are standard (Bottou et al., 2018; Dossal et al., 2024).

Proof of Proposition 1. From (10), the sequences $a_k = \gamma_k \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k}^2$ and $b_k = \gamma_k^{-1}$ verify

$$\sum_{k\geq 1} a_k < +\infty \qquad \text{and} \qquad \lim_{k\to+\infty} b_k = +\infty.$$

Hence, by Kronecker's lemma,

$$\lim_{K\to+\infty} \gamma_K \sum_{k=1}^{K} \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|_{\sigma_k}^2 = 0. \tag{36}$$

Let $\epsilon > 0$. Markov's inequality yields,

$$\mathbb{P}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|_{\sigma_{i(K)}}^2 > \epsilon\big) \leq \frac{1}{\epsilon}\mathbb{E}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|_{\sigma_{i(K)}}^2\big).$$

The above expectation is taken with respect to the stochastic iterates σ_k as well as the random choice of $i(K)$. Since these two sources of randomness are independent,

$$\mathbb{E}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|_{\sigma_{i(K)}}^2\big) = \mathbb{E}_K\mathbb{E}_{i(K)}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|_{\sigma_{i(K)}}^2\big),$$

where $\mathbb{E}_K$ denotes the expectation over the stochastic iterates $\sigma_1, \cdots, \sigma_K$. Therefore,

$$\mathbb{P}\big(\|\nabla\mathscr{F}(\sigma_{i(K)})\|_{\sigma_{i(K)}}^2 > \epsilon\big) \leq \frac{1}{\epsilon}\mathbb{E}_K\left(\frac{1}{K}\sum_{k=1}^{K}\|\nabla\mathscr{F}(\sigma_k)\|_{\sigma_k}^2\right),$$

which converges towards 0 from (36) with $\gamma_K \geq 1/K$ together with the dominated convergence theorem, the domination assumption coming from the boundedness of gradients in Proposition 7.

Proof of Proposition 2. Taking the expectation in (9), rearranging and telescoping the sum (with $-\mathscr{F}(\sigma_{K+1}) \leq 0$) yields

$$\sum_{k=0}^{K} \gamma_k \mathbb{E}\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k} \leq \mathscr{F}(\sigma_0) + \sum_{k=0}^{K} \gamma_k^2 \mathbb{E}\mathscr{F}(\sigma_k).$$

By Proposition 7, $\mathscr{F}(\sigma_k) \leq 2\mathrm{M}_2(\mu)$. The final result follows from dividing both sides of the above inequality by $\sum_{k=1}^{K} \gamma_k$.

Appendix C. Proofs of Section 4: Łojasiewicz inequalities

C.1. Proof of Proposition 4: a PL-like inequality for smooth densities

By (flat) convexity over densities equipped with the 2-norm (38),

$$\mathscr{F}(\sigma) \leq \int \mathscr{F}'[\sigma]\mathrm{d}(\sigma - \mu) \leq \int (\mathscr{F}'[\sigma] - c)\big(\frac{\mathrm{d}\sigma}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\big)\mathrm{d}\nu,$$

where the second inequality uses the notation $c = \int \mathscr{F}'[\sigma]\mathrm{d}\nu$ and the fact that $\int c(\sigma - \mu)\mathrm{d}\nu = 0$ since σ, μ are both probability distributions. Using the Cauchy-Schwarz inequality, and then the Poincaré inequality for ν,

$$\begin{aligned}\mathscr{F}(\sigma) &\leq \Big\|\frac{\mathrm{d}\sigma}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big\|_\nu \sqrt{\mathrm{Var}_\nu(\mathscr{F}'[\sigma])} \\ &\leq C_\nu \Big\|\frac{\mathrm{d}\sigma}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big\|_\nu \big\|\nabla\mathscr{F}'[\sigma]\big\|_\nu .\end{aligned}$$

Additionally, by the boundedness assumption (13), $\|\nabla\mathscr{F}'[\sigma]\|_\nu \leq \frac{1}{m}\|\nabla\mathscr{F}'[\sigma]\|_\sigma$, and $\|\frac{\mathrm{d}\sigma}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\nu \leq \|\frac{\mathrm{d}\sigma}{\mathrm{d}\nu}\|_\nu + \|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\nu \leq 2M$, so the result follows by using that $\nabla_{W_2}\mathscr{F}(\sigma) = \nabla\mathscr{F}'[\sigma]$.

C.2. Proof of Proposition 8: a PL-like inequality for Gaussian distributions

Proposition 8 (General covariances) *Assume that $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \Lambda)$, with Σ and Λ symmetric positive definite. Then,*

$$\mathscr{F}(\sigma)^2 \leq \frac{1}{2}W_2^2(\sigma, \mu)\Big(1 + \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\Big)\|\nabla_{W_2}\mathscr{F}(\sigma)\|^2_\sigma. \tag{37}$$

As recalled in Appendix A, a Kantorovich potential is solution of the dual formulation of OT. For ψ_θ the Kantorovich potential associated with the transport from σ^θ to μ^θ, let $\Psi(x) = \int \psi_\theta(x^\top\theta)d\mathcal{U}(\theta)$, so that

$$SW_2^2(\sigma, \mu) = \int \Psi d\sigma + \int\int \psi_\theta^c(y^\top\theta)\mathrm{d}\mathcal{U}(\theta)\mathrm{d}\mu(y).$$

Note that this dual formulation was recently proven for generalized sliced metrics (Kitagawa and Takatsu, 2024, Main Theorem, (6)). Then,

$$\int \Psi\mathrm{d}\sigma - \int \Psi\mathrm{d}\mu = \int \Psi\mathrm{d}\sigma - \int\int \psi_\theta(y^\top\theta)\mathrm{d}\mathcal{U}(\theta)\mathrm{d}\mu(y).$$

By definition of the c-transform, we have for all u, v that $\psi_\theta^c(u) \leq \frac{1}{2}\|u-v\|^2 - \psi_\theta(v)$, and, for $v = u$, $\psi_\theta^c(u) \leq -\psi_\theta(u)$. As a byproduct, one recovers

$$\frac{1}{2}SW_2^2(\sigma,\mu) = \int \Psi \mathrm{d}\sigma + \int \psi_\theta^c(y^\top\theta)\mathrm{d}\mathcal{U}(\theta)\mathrm{d}\mu(y) \leq \int \Psi \mathrm{d}(\sigma-\mu). \tag{38}$$

We stress that the above is a rewriting of the (flat) convexity of L with respect to the 2-norm between densities, because Ψ is the first variation of L at σ, (Cozzi and Santambrogio, 2025). Since Ψ is locally Lipschitz (Rockafellar, 1970, Theorem 10.4), a direct application of Chewi et al. (2020, Lemma 13) yields

$$\left|\int \Psi \mathrm{d}\sigma - \int \Psi \mathrm{d}\mu\right| \leq W_2(\sigma,\mu)\int_0^1 \|\nabla\Psi\|_{\rho_t}\mathrm{d}t, \tag{39}$$

where $\rho_t = ((1-t)\mathrm{Id} + tT_\sigma^\mu)_\sharp\sigma$ is the Wasserstein geodesic between σ and μ. Combining (38) and (39),

$$SW_2^2(\sigma,\mu) \leq W_2(\sigma,\mu)\int_0^1 \|\nabla\Psi\|_{\rho_t}\mathrm{d}t.$$

Taking the square and applying Jensen's inequality,

$$SW_2^4(\sigma,\mu) \leq W_2^2(\sigma,\mu)\int_0^1 \|\nabla\Psi\|_{\rho_t}^2\mathrm{d}t.$$

We stress that $\nabla\Psi(x) = \int \theta(x^\top\theta - T_{\sigma^\theta}^{\mu^\theta}(x^\top\theta))d\mathcal{U}(\theta) = (1/d)\nabla_{W_2}\mathscr{F}(\sigma)(x)$, so

$$4\mathscr{F}(\sigma)^2 \leq W_2^2(\sigma,\mu)\int_0^1 \|\nabla_{W_2}\mathscr{F}(\sigma)\|_{\rho_t}^2\mathrm{d}t. \tag{40}$$

Thus, it only remains to show that $\int_0^1 \|\nabla_{W_2}\mathscr{F}(\sigma)\|_{\rho_t}^2\mathrm{d}t \lesssim \|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2$. Under conditions on eigenvalues and the linearity of OT maps for Gaussian distributions, we proceed with the same arguments as in the proof of Chewi et al. (2020, Theorem 19). Since $\sigma^\theta = \mathcal{N}(0,\theta^\top\Sigma\theta)$ and $\mu^\theta = \mathcal{N}(0,\theta^\top\Lambda\theta)$, we have $T_{\sigma^\theta}^{\mu^\theta} : z \mapsto \tau_\theta z$ for $\tau_\theta = \sqrt{\theta^\top\Lambda\theta/\theta^\top\Sigma\theta}$. As a byproduct,

$$\nabla_{W_2}\mathscr{F}(\sigma)(x) = d\int \theta(x^\top\theta - T_{\sigma^\theta}^{\mu^\theta}(x^\top\theta))\mathrm{d}\mathcal{U}(\theta) = d\int(1-\tau_\theta)\theta\theta^\top\mathrm{d}\mathcal{U}(\theta)x = Ax,$$

for $A = d\int(1-\tau_\theta)\theta\theta^\top\mathrm{d}\mathcal{U}(\theta)$. Denote by $B = \Sigma^{-1/2}(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Sigma^{-1/2}$ such that the OT map from σ to μ verifies $T_\sigma^\mu(x) = Bx$. Then, the integration over ρ_t writes, for $X \sim \sigma$,

$$\|\nabla_{W_2}\mathscr{F}(\sigma)\|_{\rho_t}^2 = \mathbb{E}\|(1-t)AX + tABX\|^2 \leq (1-t)\mathbb{E}\|AX\|^2 + t\mathbb{E}\|ABX\|^2 \tag{41}$$

Because $BX \sim \mathcal{N}(0,\Lambda)$, one has $ABX \sim \mathcal{N}(0, A\Lambda A^\top)$ and thus $\mathbb{E}\|ABX\|^2 = \mathrm{Tr}(A\Lambda A) = \mathrm{Tr}(\Lambda A^2)$. Using the von Neumann's trace inequality (singular values coincide with eigenvalues for normal and positive matrices),

$$\mathbb{E}\|ABX\|^2 = \mathrm{Tr}(\Lambda\Sigma^{-1}\Sigma A^2) \leq \sum_{i=1}^d \lambda_i(\Lambda\Sigma^{-1})\lambda_i(\Sigma A^2) \leq \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\mathrm{Tr}(\Sigma A^2) = \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\mathbb{E}\|AX\|^2,$$

Plugging this in (41) induces

$$\int_0^1 \|\nabla_{W_2}\mathscr{F}(\sigma)\|_{\rho_t}^2\mathrm{d}t \leq \frac{1}{2}\Big(1+\frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\Big)\mathbb{E}\|AX\|^2 \leq \frac{1}{2}\Big(1+\frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\Big)\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2,$$

and the results follows by combining with (40).

C.3. Proof of Proposition 5

This section refines the PL-like inequality between Gaussian distributions with co-diagonalizable covariance matrices. Proposition 5 stems upon the following result.

Proposition 9 *Consider two centered Gaussian measures μ_Σ and μ_Λ in $\mathbb{R}^d$ with diagonal covariance matrices Σ and Λ. Assume there exists finite constants $0 < m \leq M$ such that all diagonal entries of Σ and Λ lie in $[m, M]$. Then,*

$$SW_2^2(\mu_\Sigma, \mu_\Lambda) \geq \frac{m}{Md(d+2)} W_2^2(\mu_\Sigma, \mu_\Lambda)\,. \tag{42}$$

Proof For $i \in \{1, \dots, d\}$, denote by σ_i^2 and λ_i^2 the i-th diagonal element of Σ and Λ respectively. By the closed-form solution of the Wasserstein distance of order 2 between Gaussians,

$$W_2^2(\mu_\Sigma, \mu_\Lambda) = \|\Sigma^{1/2} - \Lambda^{1/2}\|_F^2 = \sum_{i=1}^d (\sigma_i - \lambda_i)^2\,.$$

On the other hand, the Sliced-Wasserstein distance is defined as

$$SW_2^2(\mu_\Sigma, \mu_\Lambda) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[(\sqrt{\theta^\top \Sigma \theta} - \sqrt{\theta^\top \Lambda \theta})^2]\,. \tag{43}$$

For all $(x, y) \in \mathbb{R}^2$, $(\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) = x - y$. Additionally, if $0 < x, y < M$,

$$(\sqrt{x} - \sqrt{y})^2 \geq \frac{(x-y)^2}{4M}\,.$$

Since for all $i \in \{1, \dots, d\}$, σ_i^2 and λ_i^2 are bounded between m and M, so are $\theta^\top \Sigma \theta$ and $\theta^\top \Lambda \theta$. We can thus bound (43) as,

$$SW_2^2(\mu_\Sigma, \mu_\Lambda) \geq \frac{1}{4M}\, \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[(\theta^\top \Gamma \theta)^2]\,, \tag{44}$$

where $\Gamma = \Sigma - \Lambda$. Since θ is uniformly distributed on the sphere, one can show (Wiens, 1992)

$$\mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[(\theta^\top \Gamma \theta)^2] = \frac{2\mathrm{Tr}(\Gamma^2) + (\mathrm{Tr}(\Gamma))^2}{d(d+2)}\,.$$

The final result follows from $\mathrm{Tr}(\Gamma)^2 \geq 0$ and

$$\begin{aligned}
\mathrm{Tr}(\Gamma^2) &= \sum_{i=1}^d (\sigma_i^2 - \lambda_i^2)^2 = \sum_{i=1}^d (\sigma_i - \lambda_i)^2 (\sigma_i + \lambda_i)^2 \\
&\geq 4m \sum_{i=1}^d (\sigma_i - \lambda_i)^2 \\
&\geq 4m W_2^2(\mu_\Sigma, \mu_\Lambda)\,.
\end{aligned}$$

■

Remark 11 (Extension to elliptically contoured distributions) *Proposition 9 can be readily extended to the class of elliptically contoured distributions whose positive definite parameters are co-diagonalizable.*

Proof *(Proof of Proposition 5)* By Proposition 9, for co-diagonalizable covariance matrices, there exists $C_{m,d} > 0$ such that $W_2^2(\sigma,\mu) \leq SW_2^2(\sigma,\mu)C_{m,d}$. We conclude by rearranging terms in Proposition 8. ■

Appendix D. Proofs of Section 4.4: Eigenvalues control along the iterations

Objective and bottleneck. Recall that the inequality provided in (37) writes

$$\mathscr{F}(\sigma)^2 \leq \frac{1}{2}W_2^2(\sigma,\mu)\Big(1+\frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma)}\Big)\|\nabla_{W_2}\mathscr{F}(\sigma)\|_\sigma^2.$$

In order to use this inequality for convergence rates, one only needs to control eigenvalues along the iterations, as $W_2(\sigma_k,\mu)$ is bounded from Proposition 7. This is the purpose of the remaining of the section. Firstly, the following recursion for covariances of $(\sigma_k)_k$ is known to hold for Wasserstein geodesics between Gaussians, (Altschuler et al., 2021, Appendix A):

$$\Sigma_{k+1} = ((1-\gamma_k)\mathrm{Id} + \gamma_k T_{P_{k+1}})\Sigma_k((1-\gamma_k)\mathrm{Id} + \gamma_k T_{P_{k+1}}),$$

where $T_{P_{k+1}} = P_{k+1}D_kP_{k+1}$ is the matrix form of the sliced map from σ_k to μ in the directions P_{k+1} (it will be detailed in the next Proposition 10). A convenient feature is that eigenvalues can be controlled along such Wasserstein geodesics, by eigenvalues of σ_k and $T_{P_{k+1}\sharp}\sigma_k$ (Chewi et al., 2020; Altschuler et al., 2021). Nonetheless, in our particular setting, sliced maps do not necessarily push the source forward onto the target. Hence, the covariance matrix of $T_{P_{k+1}\,\sharp}\sigma_k$ is not necessarily the one of μ, and control of eigenvalues is not a direct byproduct of assumptions on μ.

Sketch. This section is structured as follows. A recursive inequality for eigenvalues of the covariance matrix of $T_{P_{k+1}\,\sharp}\sigma_k$ is given in Proposition 10. It includes randomness coming from the stochastic gradients and the choice of projection directions. The latter randomness is controlled in Proposition 11 by bounding expectations with Lemma 5, assuming that the target μ is isotropic. If instead μ has a general covariance matrix, Proposition 12 gives only a sufficient condition.

D.1. Recursive inequalities on eigenvalues

Proposition 10 *Assume that $\sigma_k = \mathcal{N}(0,\Sigma_k)$ and $\mu = \mathcal{N}(0,\Lambda)$, with Σ_k and Λ symmetric positive definite. Then, there exist directions θ_i, θ_j taken from the basis P_{k+1} such that, for $\tau_\theta = \sqrt{\theta^\top\Lambda\theta/\theta^\top\Sigma\theta}$,*

$$\sqrt{\lambda_{\min}(\Sigma_k)}\big(1+\gamma_k(\tau_{\theta_i}-1)\big) \leq \sqrt{\lambda_{\min}(\Sigma_{k+1})} \leq \sqrt{\lambda_{\max}(\Sigma_{k+1})} \leq \sqrt{\lambda_{\max}(\Sigma_k)}\big(1+\gamma_k(\tau_{\theta_j}-1)\big). \tag{45}$$

In particular, Σ_{k+1} is symmetric positive definite.

Proof The distribution σ_{k+1} corresponds to the random vector

$$(1-\gamma_k)X + \gamma_k T_{P_{k+1}}(X), \tag{46}$$

where $X \sim \mathcal{N}(0, \Sigma_k)$. Also, by definition,

$$T_{P_{k+1}}(X) = \sum_{\ell=1}^{d} \theta_\ell t_{\theta_\ell}(X^\top \theta_\ell) = \sum_{\ell=1}^{d} \tau_{\theta_\ell} \theta_\ell \theta_\ell^\top X = P_{k+1} D_k P_{k+1}^\top X,$$

where $D_k = \mathrm{diag}(\tau_{\theta_1}, \dots, \tau_{\theta_d})$ is positive definite. With these notations, $T_{P_{k+1}}(X) \sim \mathcal{N}(0, \Gamma)$ with $\Gamma = P_{k+1} D_k P_{k+1}^\top \Sigma_k P_{k+1} D_k P_{k+1}^\top$ and $T_{P_{k+1}}$ is the gradient of a convex function.

As a byproduct, the interpolate (46) belongs to the path $t \mapsto ((1-t)\mathrm{Id} + tT_{P_{k+1}})_\sharp \sigma_k$ that is a Wasserstein geodesic bridging two Gaussian distributions. The functionals $-\sqrt{\lambda_{\min}}$ and $\sqrt{\lambda_{\max}}$ have been shown to be convex along barycenters (Altschuler et al., 2021, Theorem 6), a fortiori convex along Wasserstein geodesics (Agueh and Carlier, 2011, Proposition 7.3). In other words,

$$\begin{aligned}(1-\gamma_k)\sqrt{\lambda_{\min}(\Sigma_k)} + \gamma_k \sqrt{\lambda_{\min}(\Gamma)} \le \sqrt{\lambda_{\min}(\Sigma_{k+1})} \\ \sqrt{\lambda_{\max}(\Sigma_{k+1})} \le (1-\gamma_k)\sqrt{\lambda_{\max}(\Sigma_k)} + \gamma_k \sqrt{\lambda_{\max}(\Gamma)}. \end{aligned} \tag{47}$$

Hence, it remains to control eigenvalues of Γ. On the one hand, $\overline{\Sigma} = P_{k+1}^\top \Sigma_k P_{k+1}$ and Σ_k have the same eigenvalues, by orthonormality of P_{k+1}[5]. On the other hand, Γ has the same eigenvalues as $D\overline{\Sigma}D_k$ from the same argument. Also, D is non singular because Σ_k and Λ have positive eigenvalues, hence $\tau_{\theta_\ell} > 0$ for all $\ell = 1, \cdots, d$. Then, a direct application of Ostrowski's Theorem (Ostrowski, 1959) entails that

$$\lambda_i(\Gamma) = \lambda_i(D_k \overline{\Sigma} D_k) = \beta_i \lambda_i(\Sigma_k), \tag{48}$$

with

$$\min_j \frac{\theta_j^\top \Lambda \theta_j}{\theta_j^\top \Sigma_k \theta_j} \le \beta_i \le \max_j \frac{\theta_j^\top \Lambda \theta_j}{\theta_j^\top \Sigma_k \theta_j}.$$

Thus, the result follows by combining (47) and (48). ■

D.2. A bound in expectation for isotropic target

Proposition 11 gives a deterministic upper bound on eigenvalues of (Σ_k), and a lower bound in expectation. It requires bounds on p-moments of $\theta^\top \Sigma_k \theta - 1$, that are provided just after in Lemma 5.

Proposition 11 *Assume that $\sigma_0 = \mathcal{N}(0, \Sigma_0)$, with $\Sigma_0 \in \mathbb{R}^{d\times d}$ symmetric, positive-definite, and $\mu = \mathcal{N}(0, \mathbf{I}_d)$. Then, for any $k \ge 1$, the IDT iterates remain Gaussian, $\sigma_k = \mathcal{N}(0, \Sigma_k)$ with*

$$\mathbb{E}[1/\lambda_{\min}(\Sigma_k)] \le \mathbb{E}[1/\lambda_{\min}(\Sigma_1)] \tag{49}$$

$$\forall p \in \mathbb{N}^*, \quad \mathbb{E}[1/\lambda_{\min}(\Sigma_k)^p] \le \mathbb{E}[1/\lambda_{\min}(\Sigma_1)^p] \prod_{l=1}^{k} (1 + B_p \gamma_l^2)\,. \tag{50}$$

where $B_p > 0$. Note that $\prod_{l=1}^{\infty}(1 + B_p\gamma_l^2)$ is finite for any step-sizes sequence $(\gamma_k)_k$ satisfying (5).

5. Eigenvectors of $\overline{\Sigma}$ are of the form $P_{k+1}^\top u$ for u an eigenvector of Σ_k. Indeed, $(P_{k+1}^\top u)^\top P_{k+1}^\top \Sigma_k P_{k+1}(P_{k+1}^\top u) = u^\top \Sigma_k u$ which equals an eigenvalue of Σ_k.

Proof A direct byproduct of Proposition 7 is that $\lambda_{\min}(\Sigma_k) \leq \mathrm{Tr}(\Sigma_k) \leq \mathrm{Tr}(\mathrm{Id}) = d$. We now focus on showing (49).

From (45), there exists $1 \leq i \leq d$ such that $\sqrt{\lambda_{\min}(\Sigma_k)}\big(1 + \gamma_k(\tau_{\theta_i} - 1)\big) \leq \sqrt{\lambda_{\min}(\Sigma_{k+1})}$. Taking the inverse and using that the harmonic mean is always smaller than the arithmetic mean,

$$(\lambda_{\min}(\Sigma_{k+1}))^{-1/2} \leq (\lambda_{\min}(\Sigma_k))^{-1/2}\big(1 - \gamma_k + \gamma_k \tau_{\theta_i}^{-1}\big).$$

Here, everything is positive due to the positivity of all $(\tau_{\theta_i})_i$, so that taking the square and applying Jensen's inequality yields

$$(\lambda_{\min}(\Sigma_{k+1}))^{-1} \leq (\lambda_{\min}(\Sigma_k))^{-1}\big(1 - \gamma_k + \gamma_k \theta_i^\top \Sigma_k \theta_i\big). \tag{51}$$

Recall that θ_i belongs to the random basis P_{k+1}, whose distribution is independent from the σ-field $\mathcal{A}_k$ generated by $P_1, \ldots, P_k$. Also, Σ_k is measurable with respect to $\mathcal{A}_k$. Hence, taking the conditional expectation in (51) yields

$$\mathbb{E}[(\lambda_{\min}(\Sigma_{k+1}))^{-1}|\mathcal{A}_k] \leq (\lambda_{\min}(\Sigma_k))^{-1}\big(1 + \gamma_k \mathbb{E}\Big[\theta_i^\top \Sigma_k \theta_i - 1|\mathcal{A}_k\Big]\big).$$

By independence between the distribution of θ_j and $\mathcal{A}_k$, and by the $\mathcal{A}_k$-measurability of Σ_k,

$$\mathbb{E}\big[\theta_i^\top \Sigma_k \theta_i - 1|\mathcal{A}_k\big] = \mathbb{E}_\theta\big[\theta^\top \Sigma_k \theta\big] - 1 = \frac{1}{d}\mathrm{Tr}(\Sigma_k) - 1.$$

However, moments are bounded along iterations from Proposition 7, so $\mathrm{Tr}(\Sigma_k) \leq \mathrm{Tr}(\mathrm{Id}) = d$. Combining this with the two equations above induces

$$\mathbb{E}[(\lambda_{\min}(\Sigma_{k+1}))^{-1}|\mathcal{A}_k] \leq (\lambda_{\min}(\Sigma_k))^{-1},$$

and (49) follows by induction.

Now, fix $p \geq 2$. Taking the power p and applying Jensen's inequality in (51) induces

$$(\lambda_{\min}(\Sigma_{k+1}))^{-p} \leq (\lambda_{\min}(\Sigma_k))^{-p}\big(1 - \gamma_k + \gamma_k \theta_i^\top \Sigma_k \theta_i\big)^p. \tag{52}$$

By the binomial theorem, for $Z_{k,i} = \theta_i^\top \Sigma_k \theta_i - 1$,

$$\big(1 + \gamma_k Z_{k,i}\big)^p = 1 + p\gamma_k Z_{k,i} + \sum_{r=2}^{p} \binom{p}{r} \gamma_k^r Z_{k,i}^r.$$

Taking the expectation with respect to $\mathcal{A}_k$, and using upper-bounds from Lemma 5,

$$\mathbb{E}[\big(1 + \gamma_k Z_{k,i}\big)^p|\mathcal{A}_k] \leq 1 + \gamma_k^2 \sum_{r=2}^{p} \binom{p}{r} \gamma_k^{r-2} \mathbb{E}(Z_{k,i}^r|\mathcal{A}_k) < +\infty.$$

Plugging this in (52), and reasoning by induction, it exists $B > 0$ such that the desired result holds. ■

Lemma 5 *Let $A \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix verifying $\mathrm{Tr}(A) \leq d$. For θ uniformly distributed over the unit sphere,*

$$\mathbb{E}_\theta(\theta^\top A\theta - 1) \leq 0,$$

and, for all $p \geq 2$,

$$\mathbb{E}_\theta[(\theta^\top A\theta - 1)^p] \leq 1 + \sum_{r=1}^{p} \binom{p}{r} \lambda_{\max}(A)^{r-1}(-1)^r < +\infty.$$

Proof The first point is a byproduct of $\mathbb{E}_\theta(\theta^\top A\theta - 1) = \mathrm{Tr}(A)/d - 1$. From the fact that $\mathrm{Tr}(a) = a$ if $a \in \mathbb{R}$, and the cyclic property of Tr,

$$\mathbb{E}_\theta[(\theta^\top A\theta)^2] = \mathbb{E}_\theta Tr[\theta^\top A\theta\theta^\top A\theta] \leq \mathbb{E}_\theta Tr[\theta\theta^\top A\theta\theta^\top A].$$

Because A and $\theta\theta^\top$ are positive semi-definite, the von Neumann's trace inequality implies

$$\mathbb{E}_\theta[(\theta^\top A\theta)^2] \leq \mathbb{E}_\theta\big(\lambda_{\max}(\theta\theta^\top)Tr[A\theta\theta^\top A]\big) = \mathbb{E}_\theta\big(\lambda_{\max}(\theta\theta^\top)Tr[A^2\theta\theta^\top]\big).$$

By linearity of $\mathbb{E}_\theta$ and Tr, together with $\lambda_{\max}(\theta\theta^\top) \leq 1$ and $\mathbb{E}_\theta[\theta\theta^\top] = \mathrm{Id}/d$,

$$\mathbb{E}_\theta[(\theta^\top A\theta)^2] \leq Tr[A^2\mathbb{E}_\theta(\theta\theta^\top)] = \frac{\mathrm{Tr}(A^2)}{d}.$$

Using again the von Neumann's trace inequality, $\mathrm{Tr}(A^2) \leq \lambda_{\max}(A)\mathrm{Tr}(A)$, and $\mathrm{Tr}(A) \leq d$, so $\mathrm{Tr}(A^2)/d \leq \lambda_{\max}(A)$ which proves the first point:

$$\mathbb{E}_\theta((\theta^\top A\theta - 1)^2) = \mathbb{E}_\theta[(\theta^\top A\theta)^2 + 1 - 2\theta^\top A\theta] \leq \lambda_{\max}(A) + 1.$$

With the same arguments as above, one can deduce that, for all $p \geq 1$,

$$\mathbb{E}_\theta[(\theta^\top A\theta)^p] \leq \frac{\mathrm{Tr}(A^p)}{d} \leq \lambda_{\max}(A)^{p-1}.$$

Thus, the last claims follows by the binomial theorem,

$$\mathbb{E}_\theta[(1 - \theta^\top A\theta)^p] = \mathbb{E}_\theta \sum_{r=0}^{p} \binom{p}{r} (\theta^\top A\theta)^r(-1)^r \leq 1 + \sum_{r=1}^{p} \binom{p}{r} \lambda_{\max}(A)^{r-1}(-1)^r.$$

■

D.3. A sufficient condition under arbitrary covariance

Let $\mu = \mathcal{N}(0, \Lambda)$ for a general covariance matrix Λ.

Proposition 12 *For all $p \geq 1$, a sufficient condition for the existence of a finite constant $C_p > 0$ such that*

$$\sup_{k\in\mathbb{N}} \mathbb{E}[(\lambda_{\min}(\Sigma_k))^{-p}] \leq C_p$$

is the following,

$$\mathbb{E}\left(\sum_{k\geq 0}^{\infty} \gamma_k \mathbb{E}_\theta\Big[\Big(\frac{\theta^\top\Sigma_k\theta}{\theta^\top\Lambda\theta}\Big)^p - 1\Big]\right) < +\infty.$$

Proof As a byproduct of Proposition 10, and proceeding as in the beginning of Proposition 11, the following counterpart of (51) holds,

$$(\lambda_{\min}(\Sigma_{k+1}))^{-1} \leq (\lambda_{\min}(\Sigma_k))^{-1}\big(1-\gamma_k+\gamma_k\frac{\theta_i^\top\Sigma_k\theta_i}{\theta_i^\top\Lambda\theta_i}\big).$$

Fix $p \geq 1$, and apply the power p and Jensen's inequality to obtain that

$$(\lambda_{\min}(\Sigma_{k+1}))^{-p} \leq (\lambda_{\min}(\Sigma_k))^{-p}\Big(1-\gamma_k+\gamma_k\Big(\frac{\theta_i^\top\Sigma_k\theta_i}{\theta_i^\top\Lambda\theta_i}\Big)^p\Big).$$

Taking the conditional expectation,

$$\mathbb{E}[(\lambda_{\min}(\Sigma_{k+1}))^{-p}|\mathcal{A}_k] \leq (\lambda_{\min}(\Sigma_k))^{-p}\Big(1+\gamma_k\mathbb{E}_\theta\Big[\Big(\frac{\theta^\top\Sigma_k\theta}{\theta^\top\Lambda\theta}\Big)^p-1\Big]\Big).$$

Taking the expectation and reasoning by induction, we deduce that

$$\mathbb{E}[(\lambda_{\min}(\Sigma_{k+1}))^{-p}] \leq \mathbb{E}[(\lambda_{\min}(\Sigma_0))^{-1}] + \mathbb{E}\left(\sum_{j=0}^{k}\gamma_k\mathbb{E}_\theta\Big[\Big(\frac{\theta^\top\Sigma_k\theta}{\theta^\top\Lambda\theta}\Big)^p-1\Big]\right).$$

Thus: $\forall p \geq 1, \exists C_p > 0,\ \sup_{k\in\mathbb{N}}\mathbb{E}[(\lambda_{\min}(\Sigma_k))^{-p}] \leq C_p$. ■

Appendix E. Proof of our main result: Theorem 2

E.1. Proof of Proposition 3

Recall that Robbins and Siegmund (1971) implies $(\mathscr{F}(\sigma_k))_{k\geq 0}$ converges almost surely to a finite random variable, as a direct byproduct of (9). Hence, one only needs to show that the limit random variable is zero $a.s.$ For this, two assumptions are needed in Li et al. (2023, Theorem 2), with standard arguments (Duflo, 1996): (i) $(\sigma_k)_{k\geq 1}$ remains in a compact subset K of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ and (ii) $\nabla\mathscr{F}(\sigma) = 0 \Longrightarrow \sigma = \mu$ for $\sigma \in K$. Under continuity and compact supports, Li et al. (2023) show (i), while (ii) holds for absolutely continuous measures from (Bonnotte, 2013, Lemma 5.7.2). Taken together, these two facts imply almost-sure convergence.

To extend this result to Gaussian measures, we verify (i) and (ii) hereafter. By Li et al. (2023, Propositions 4 and 5), the iterates σ_k remain in a compact set of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ if $\sigma, \mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and have finite third-order moments. This is verified for Gaussian distributions, hence there is a set K such that (i) holds. In addition, by Proposition 10, each iterate σ_k remains Gaussian with strictly positive definite covariance Σ_k for every finite k. Therefore, regarding (ii), Proposition 8 gives that

$$\mathscr{F}(\sigma_k)^2 \leq 2\mathrm{M}_2(\mu)\Big(1+\frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Sigma_k)}\Big)\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k},$$

where we also use that $W_2^2(\sigma_k,\mu) \leq 4\mathrm{M}_2(\mu)$ from Proposition 7. Consequently, (ii) is verified under general Gaussian covariances, and the desired result follows.

E.2. Proof of Theorem 2

By Proposition 5, the following PL condition holds for any $k \geq 1$,

$$\mathscr{F}(\sigma_k) \leq \frac{C_{k,d}}{2}\Big(1 + \frac{1}{\lambda_{\min}(\Sigma_k)}\Big)\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}\,, \tag{53}$$

with $C_{k,d} = d(d+2)M_k/m_k$, $M_k = \max(\lambda_{\max}(\Sigma_k), 1)$ and $m_k = \min(\lambda_{\min}(\Sigma_k), 1)$. By Proposition 11, $M_k \leq d$, thus $C_{k,d} \leq d^2(d+2)/m_k$. Additionally, by Proposition 7, we have $\mathrm{Tr}(\Sigma_k) \leq \mathrm{Tr}(\Lambda)$ where Λ denotes the covariance matrix of the target Gaussian. A contradiction argument then implies $\lambda_{\min}(\Sigma_k) \leq \lambda_{\max}(\Lambda)$, and in the special case $\Lambda = \mathbf{I}_d$, this gives $\lambda_{\min}(\Sigma_k) \leq 1$. Therefore, $C_{k,d} \leq d^2(d+2)/\lambda_{\min}(\Sigma_k)$ (since $m_k = \lambda_{\min}(\Sigma_k)$), and $1 + 1/\lambda_{\min}(\Sigma_k) \leq 2/\lambda_{\min}(\Sigma_k)$. Therefore, (53) entails that

$$\mathscr{F}(\sigma_k) \leq \frac{d^2(d+2)}{\lambda_{\min}(\Sigma_k)^2}\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}\,. \tag{54}$$

Denote by $B_k = d^2(d+2)/\lambda_{\min}(\Sigma_k)^2$. Proposition 11 gives us that all the moments of B_k are finite: $\sup_k \mathbb{E}[B_k^p] \leq c_p < +\infty$ for all $p \in \mathbb{N}^*$. In other words, the expected PL inequality in Assumption A for $\tau = 1$ holds along the iterates σ_k. Thus, the result is a byproduct of Theorem 3.

E.3. Proof of Theorem 3

Random events and main recursion. Since $(B_k)_{k\geq 1}$ is a sequence of random variables (Assumption A), we condition the analysis on the event $G_k = \{B_k \leq 1/g_k\}$ to apply the PL inequality. This is done by introducing a sequence of positive numbers $(g_k)_{k\geq 1}$ with $\lim_{k\to+\infty} g_k = 0$, so that $\mathbb{1}_{G_k}$ converges to 1 almost surely. Denote by G_k^c the complementary event: $G_k^c = \{B_k > 1/g_k\}$.

Expected PL inequality. We begin with the rates obtained under Assumption A with $\tau = 1$. Using the descent lemma (9) and the PL inequality of Assumption A on the event G_k,

$$\begin{aligned}\mathbb{E}[\mathscr{F}(\sigma_{k+1})|\mathcal{A}_k] &\leq (1+\gamma_k^2)\mathscr{F}(\sigma_k) - \gamma_k\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}(\mathbb{1}_{G_k} + \mathbb{1}_{G_k^c}),\\ &\leq (1+\gamma_k^2)\mathscr{F}(\sigma_k) - \gamma_k g_k\mathscr{F}(\sigma_k)\mathbb{1}_{G_k} - \gamma_k\|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k}\mathbb{1}_{G_k^c}.\end{aligned}$$

Now, we can plug the decomposition (12), that is $2\mathcal{F}(\sigma_k) - \|\nabla_{W_2}\mathscr{F}(\sigma_k)\|^2_{\sigma_k} = E_k$ for $E_k = \mathbb{E}_P\big[\|\overline{T}_{\sigma_k} - T_{\sigma_k,P}\|^2_{\sigma_k}\big]$. This implies

$$\begin{aligned}\mathbb{E}[\mathscr{F}(\sigma_{k+1})|\mathcal{A}_k] &\leq (1+\gamma_k^2)\mathscr{F}(\sigma_k) - \gamma_k g_k\mathscr{F}(\sigma_k)\mathbb{1}_{G_k} - \gamma_k 2\mathscr{F}(\sigma_k)\mathbb{1}_{G_k^c} + \gamma_k E_k\mathbb{1}_{G_k^c}, &(55)\\ &\leq (1+\gamma_k^2-\gamma_k g_k)\mathscr{F}(\sigma_k)\mathbb{1}_{G_k} + (1-\gamma_k)^2\mathscr{F}(\sigma_k)\mathbb{1}_{G_k^c} + \gamma_k E_k\mathbb{1}_{G_k^c}. &(56)\end{aligned}$$

The first term holds under the event G_k, where the PL inequality holds. Depending on the choice of g_k (which, for now, only needs to converge to zero), this will result in an exponential rate governed by $\exp(\sum_{j=0}^k \gamma_k^2 - \gamma_k g_k)$ since $\prod_i(1+a_i) \leq \sum_i \exp(a_i)$.

On the event G_k^c, we obtained two terms from the decomposition (12). One first recovers $\mathscr{F}(\sigma_k)(1-\gamma_k)^2$, which, if it were alone, would also imply an exponential rate. The final rate will thus be governed by $\gamma_k E_k\mathbb{1}_{G_k^c}$.

In the absence of any convergence result for E_k, we observe that $E_k \leq 2\mathcal{F}(\sigma_k) \leq 4\mathrm{M}_2(\mu)$ from Proposition 7. Therefore, $\gamma_k E_k \mathbb{1}_{G_k^c} \leq \gamma_k 4\mathrm{M}_2(\mu)\mathbb{1}_{G_k^c}$, and the final rate will be governed solely by $\mathbb{1}_{G_k^c}$. Combining this with the recursion (56) and using that $-2\gamma_k \leq -\gamma_k g_k$, we obtain

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq (1 + \gamma_k^2 - \gamma_k g_k)\mathbb{E}[\mathscr{F}(\sigma_k)] + 4\mathrm{M}_2(\mu)\gamma_k \mathbb{E}[\mathbb{1}_{G_k^c}]$$

after taking the expectation. By Markov's inequality,

$$\forall p \in \mathbb{N}^*, \quad \mathbb{P}(G_k^c) = \mathbb{P}(B_k > 1/g_k) \leq g_k^p \mathbb{E}[B_k^p] \leq c_p g_k^p. \tag{57}$$

Here, p can be chosen freely. Choosing a larger value of p improves the decay of the factor g_k^p, at the cost of increasing the constant c_p. This yields the main recursion: for any $p \in \mathbb{N}^*$,

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq (1 + \gamma_k^2 - \gamma_k g_k)\mathbb{E}[\mathscr{F}(\sigma_k)] + 4\mathrm{M}_2(\mu)c_p\gamma_k g_k^p. \tag{58}$$

It only remains to choose (g_k) and p. For two different choices, we obtain two different rates, whose optimality depends on α.

Case 1: $0 < \alpha < 2/3$**.** For $\gamma_k = 1/(k+1)^\alpha$, let $g_k = 1/(k+1)^\epsilon$ so that the objective rephrases as choosing ϵ and p. For any $k \geq 0$, one has for $\epsilon < \alpha$,

$$\gamma_k^2 - \gamma_k g_k = \frac{1}{(k+1)^{\alpha+\epsilon}}\Big(\frac{1}{(k+1)^{\alpha-\epsilon}} - 1\Big) \leq \frac{1}{(k+1)^{\alpha+\epsilon}}\Big(\frac{1}{2^{\alpha-\epsilon}} - 1\Big).$$

In other words, $\gamma_k^2 - \gamma_k g_k = -a/(k+1)^{\alpha+\epsilon}$ for $a = 1 - 1/2^{\alpha-\epsilon} \in (0,1)$. Thus, (58) rewrites

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq (1 - \frac{a}{(k+1)^{\alpha+\epsilon}})\mathbb{E}[\mathscr{F}(\sigma_k)] + 4\mathrm{M}_2(\mu)c_p\frac{1}{(k+1)^{\alpha+\epsilon p}}. \tag{59}$$

The desired rate follows from Bercu and Bigot (2021, Lemma A.3), which is a variant of Chung's Lemma (Chung, 1954) (see also Moulines and Bach (2011, Theorem 1)). To allow for $\alpha < 1/2$, we rewrite Bercu and Bigot (2021, Lemma A.3) in Lemma 6. This gives the existence of a constant $C > 0$ such that

$$\mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{C}{k^{\epsilon(p-1)}}, \tag{60}$$

under $\epsilon < \alpha$ and conditions that rephrase as $\alpha + \epsilon < 1 < \alpha + \epsilon p$. Equivalently, the above needs

$$\epsilon < \min(\alpha, 1-\alpha) \qquad \text{and} \qquad 1 - \alpha < \epsilon p.$$

For all $0 < \epsilon < \min(\alpha, 1-\alpha)$, there is $p \geq 1$ that meets this condition. Now, $1 - \alpha - \epsilon < \epsilon(p-1)$ yields

$$\forall \epsilon < \min(\alpha, 1-\alpha): \qquad \mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{C}{k^{\epsilon(p-1)}} \leq \frac{C}{k^{1-\alpha-\epsilon}}.$$

In particular, this is true for $0 < \alpha < 1$, even if $0 < \alpha < 1/2$ does not satsify the Robbins-Monro conditions.

Case 2: $2/3 < \alpha < 1$. We now turn to the second rate, that is faster in the regime $2/3 < \alpha < 1$. One can use Proposition 7 to bound $\gamma_k^2 \mathbb{E}[\mathscr{F}(\sigma_k)]$ by $2\mathrm{M}_2(\mu)\gamma_k^2$, hence (58) becomes

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq (1 - \gamma_k g_k)\mathbb{E}[\mathscr{F}(\sigma_k)] + 2\mathrm{M}_2(\mu)\gamma_k^2 + 4\mathrm{M}_2(\mu)c_p\gamma_k g_k^p. \tag{61}$$

By choosing $\gamma_k = 1/(k+1)^\alpha$ and $g_k = 1/(k+1)^{1-\alpha}$, we have $\gamma_k g_k = 1/(k+1)$ and $g_k^p = 1/(k+1)^{(1-\alpha)p}$. Therefore, $g_k^p \leq \gamma_k$ as soon as one chooses $p \geq \alpha/(1-\alpha)$. In this case, (61) becomes, for $C = 2\mathrm{M}_2(\mu)(1 + 2c_p)$,

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq \Big(1 - \frac{1}{k+1}\Big)\mathbb{E}[\mathscr{F}(\sigma_k)] + \frac{C}{(k+1)^{2\alpha}}\,.$$

The desired rate follows directly from Chung's Lemma (Chung, 1954):

$$\mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{C}{k^{2\alpha-1}}\,.$$

Remark 12 (Comparison of constants) *When $\alpha < 1/2$, $\min(\alpha, 1-\alpha) = \alpha$, and one needs $p > (1-\alpha)/\epsilon > (1-\alpha)/\alpha$. When $\alpha > 2/3$, the second rates requires $p \leq \alpha/(1-\alpha)$. Thus, when $\alpha \to 0$ or when $\alpha \to 1$, the order of magnitude of p is the same, and a fortiori the constant c_p. Hence, constants in both rates are similar, and this does not explain the faster convergence observed when α is almost 0 in practice.*

Expected PL-like inequality. We now turn to show the second result. Using the descent lemma (9), Assumption A with $\tau = 2$ would imply instead

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})|\mathcal{A}_k] \leq \mathscr{F}(\sigma_k)\mathbb{1}_{G_k} - \gamma_k g_k \mathscr{F}(\sigma_k)^2 \mathbb{1}_{G_k} + \mathscr{F}(\sigma_k)\mathbb{1}_{G_k^c} + \mathscr{F}(\sigma_k)\gamma_k^2, \tag{62}$$

where, in the last term, we also do not bound $\mathscr{F}(\sigma_k)$ by $2\mathrm{M}_2(\mu)$. After taking the expectation, Markov inequality can be applied as in (57) to deduce from (62) that

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq \mathbb{E}[\big(\mathscr{F}(\sigma_k) - \gamma_k g_k \mathscr{F}(\sigma_k)^2\big)\mathbb{1}_{G_k}] + 2\mathrm{M}_2(\mu)\sqrt{c_p}\, g_k^{p/2} + \mathbb{E}[\mathscr{F}(\sigma_k)]\gamma_k^2. \tag{63}$$

To remove $\mathbb{1}_{G_k}$ above, note that $\gamma_k g_k \mathscr{F}(\sigma_k)^2 \leq (2\mathrm{M}_2(\mu))^{-1}\mathscr{F}(\sigma_k)^2 \leq \mathscr{F}(\sigma_k)$, as soon as $\gamma_k g_k \leq (2\mathrm{M}_2(\mu))^{-1}$. Because g_k is a flexible choice, this just means that k needs to be large enough. For such k, we deduce $\mathbb{E}[\big(\mathscr{F}(\sigma_k) - \gamma_k g_k \mathscr{F}(\sigma_k)^2\big)\mathbb{1}_{G_k^c}] \geq 0$. Adding this to (63) gives

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq \mathbb{E}[\mathscr{F}(\sigma_k)](1 + \gamma_k^2) - \gamma_k g_k \mathbb{E}[\mathscr{F}(\sigma_k)]^2 + 2\mathrm{M}_2(\mu)\sqrt{c_p}\, g_k^{p/2}, \tag{64}$$

where we also use that $\mathbb{E}[\mathscr{F}(\sigma_k)]^2 \leq \mathbb{E}[\mathscr{F}(\sigma_k)^2]$ by Jensen's inequality. Now, all that remains is to play around with the constants to obtain the recursion necessary for an extended Chung's lemma.

Denote by $C = 2\mathrm{M}_2(\mu)\sqrt{c_p}$ and fix $\gamma = C^{3/2}$. Let $\gamma_k = 1/(k+\gamma)^\alpha$ with $1/2 < \alpha < 2/3$. Let $g_k = 1/(k+\gamma)^{2-3\alpha}$ hence

$$\gamma_k g_k = \frac{1}{(k+\gamma)^{2-2\alpha}} \qquad \text{and} \qquad g_k^{p/2} = \frac{1}{(k+\gamma)^{p(1-3\alpha/2)}}.$$

This leads to $g_k^{p/2} \leq 1/(k+\gamma)^{2\alpha}$ if $p \geq 2\alpha/(2-3\alpha)$, hence (64) rewrites

$$\mathbb{E}[\mathscr{F}(\sigma_{k+1})] \leq \mathbb{E}[\mathscr{F}(\sigma_k)]\big(1 + \frac{1}{(k+\gamma)^{2\alpha}}\big) - \frac{1}{(k+\gamma)^{2-2\alpha}}\mathbb{E}[\mathscr{F}(\sigma_k)]^2 + \frac{C}{(k+\gamma)^{2\alpha}}. \tag{65}$$

Recall that this holds as soon as $\gamma_k g_k \leq (2\mathrm{M}_2(\mu))^{-1}$, which is equivalent to $(k+\gamma)^{2-2\alpha} \geq 2\mathrm{M}_2(\mu)$. But since $\alpha < 2/3$, $(k+\gamma)^{2-2\alpha} \geq (k+\gamma)^{2/3} \geq \gamma^{2/3} \geq C \geq 2\mathrm{M}_2(\mu)$. So the recursion (65) holds for all $k \geq 0$.

We stress that this relates to an extension of Chung's Lemma in the case of a PL-type inequality with $\tau = 2$. Moulines and Bach (2011, Theorem 4) deals with a similar recursion, and Jiang et al. (2024, Lemma 19) generalizes this in several ways. Thus, it only remains to verify that (65) fulfills the correct requirements.

To stick to the notations of Jiang et al. (2024, Lemma 19), we introduce $y_k = \mathbb{E}[\mathscr{F}(\sigma_k)]$, $a_k = 1/(k+\gamma)^{2-2\alpha}$, $\ell_1 = \ell_2 = 1$, $\ell_3 = C$, $\tau = 2\alpha/(2-2\alpha)$, so that (65) rewrites, for all $k \geq 0$,

$$y_{k+1} \leq (1 + \ell_1 a_k^\tau) y_k - \ell_2 a_k y_k^2 + \ell_3 a_k^\tau.$$

This is exactly the recursion in Jiang et al. (2024, Lemma 19), and our parameters lead to their statement (b). Again, with their notations, $\zeta = C^{1/2}$, $u_2 = 2\alpha - 1$, $p = \rho$, which fulfills the requirements

$$1 \geq \big(\frac{2u_2}{\zeta}\big)^\rho \quad \text{and} \quad \gamma \geq \max\{(\frac{1}{\zeta})^{1/u_2}, \zeta\} = \zeta,$$

thus leading to

$$y_{k+1} \leq 4C^{1/2}(k+1+\gamma)^{-u_2} + y_0\big(\gamma^{-1}(k+1+\gamma)\big)^{-\zeta}.$$

This is the desired result, as it rewrites

$$\mathbb{E}[\mathscr{F}(\sigma_k)] \leq \frac{4C^{1/2}}{(k+\gamma)^{2\alpha-1}} + \frac{\mathbb{E}[\mathscr{F}(\sigma_0)]}{\big(\gamma^{-1}(k+\gamma)\big)^{\sqrt{C}}},$$

and the second term in the above is faster than the first one.

Auxiliary lemma. The next lemma is taken from Bercu and Bigot (2021, Lemma A.3). We just verify that the proof holds without the assumptions $\beta < 2$ and $\beta \leq 2\alpha$.

Lemma 6 *Let Z_k be a sequence of positive numbers satisfying, for all $k \geq 0$,*

$$Z_{k+1} \leq \big(1 - \frac{a}{(k+1)^\alpha}\big) Z_k + \frac{b}{(k+1)^\beta},$$

where a, b, α, β are positive constants satisfying $a \leq 1$ and $\alpha < 1 < \beta$. Then, there exists a positive constant C such that

$$Z_k \leq \frac{C}{k^{\beta-\alpha}}.$$

Proof Proceeding as in Bercu and Bigot (2021, Lemma A.3), with $0 \leq a \leq 1$ and $0 \leq \alpha < 1$,

$$Z_k \leq \exp(\eta(1-k^{1-\alpha}))Z_0 + b\sum_{\ell=1}^{k} P_{\ell+1}^k \frac{1}{\ell^\beta} \tag{66}$$

where $\eta = a/(1-\alpha)$ and

$$P_\ell^k = \prod_{i=\ell}^{k}(1 - \frac{a}{i^\alpha}),$$

with the convention that $P^k_{k+1} = 1$. For some integer m such that $2k \leq 4m \leq 3k$,

$$\sum_{\ell=1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta} \leq P^k_{m+1} \sum_{\ell=1}^{m} \frac{1}{\ell^\beta} + \sum_{\ell=m+1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta}, \tag{67}$$

since $P^k_{m+1} \geq P^k_{\ell+1}$ for all $\ell \leq m$. Besides, for $\xi = 1 - (3/4)^{1-\alpha}$,

$$P^k_{m+1} \leq \exp(-a \sum_{i=m+1}^{k} \frac{1}{i^\alpha}) \leq \exp(\eta(2 - \alpha - \xi k^{1-\alpha})),$$

where the first inequality uses $1 - x \leq e^{-x}$ and the second uses that $m \leq (3/4)k$ and

$$\sum_{i=m+1}^{k} \frac{1}{i^\alpha} \geq \frac{1}{1-\alpha}(k^{1-\alpha} - m^{1-\alpha}) - \frac{2-\alpha}{1-\alpha} \geq \frac{\xi k^{1-\alpha} - (2-\alpha)}{1-\alpha}.$$

Plugging this in (67) together with the bound $\sum_{\ell=1}^{m} \frac{1}{\ell^\beta} \leq 1 + \frac{1}{\beta - 1}$ that holds since $\beta > 1$,

$$\sum_{\ell=1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta} \leq \exp(\eta(2 - \alpha - \xi k^{1-\alpha}))(1 + \frac{1}{\beta - 1}) + \sum_{\ell=m+1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta}. \tag{68}$$

Regarding the right term above, remark that

$$P^k_{\ell+1} - P^k_\ell = \frac{a}{\ell^\alpha} P^k_{\ell+1}.$$

Therefore,

$$\sum_{\ell=m+1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta} = \frac{1}{a} \sum_{\ell=m+1}^{k} (P^k_{\ell+1} - P^k_\ell) \frac{1}{\ell^{\beta-\alpha}},$$

where we recall that $\beta > \alpha$. Because $\ell \geq m \geq k/2$,

$$\sum_{\ell=m+1}^{k} P^k_{\ell+1} \frac{1}{\ell^\beta} \leq \frac{2^{\beta-\alpha}}{a k^{\beta-\alpha}} \sum_{\ell=m+1}^{k} P^k_{\ell+1} - P^k_\ell \leq \frac{2^{\beta-\alpha}}{a k^{\beta-\alpha}} \big(P^k_{k+1} - P^k_{m+1}\big) \leq \frac{2^{\beta-\alpha}}{a k^{\beta-\alpha}}.$$

Combining this with (68) and the bound (66) gives the final recursion, and the desired rate,

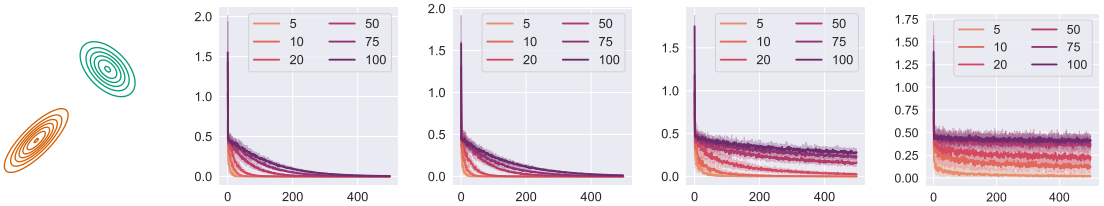
$$Z_k \leq \exp(\eta(1 - k^{1-\alpha})) Z_0 + b \exp(\eta(2 - \alpha - \xi k^{1-\alpha}))(1 + \frac{1}{\beta - 1}) + \frac{b}{a} \frac{2^{\beta-\alpha}}{k^{\beta-\alpha}}.$$

■

Appendix F. Additional Numerical Experiments

F.1. Continuous setting with explicit updates

In Figure 5, we extend the experiment of Figure 1 by considering a non-isotropic target distribution $\mu = \mathcal{N}(0, \Lambda)$, where Λ is a diagonal matrix with entries drawn from a Gaussian distribution of mean 10 and variance 1 (negative values are discarded). Conclusions are similar in this general-covariance setting, where our analysis provide convergence rates only up to the condition (21).

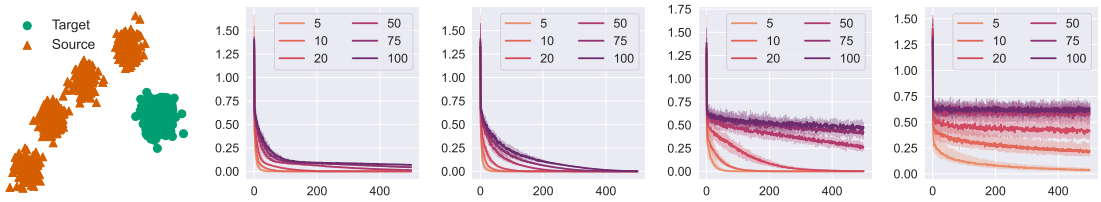


(a) Setting (b) $\alpha = 0$ (c) $\alpha = 0.1$ (d) $\alpha = 0.51$ (e) $\alpha = 0.9$

Figure 5: Evolution of $SW_2^2(\sigma_k, \mu)$ when $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \Lambda)$

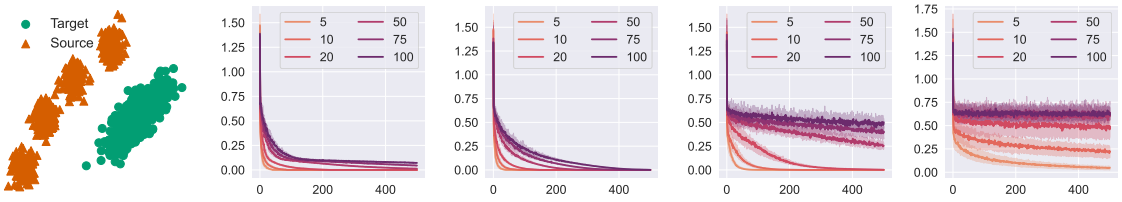
F.2. Discrete source and target

We also complement Figure 4 with Figure 6 and Figure 7 on empirical distributions sampled with $n = 500$ observations. The source is drawn from a mixture of Gaussians. The target is a Gaussian distribution, either with isotropic or non-isotropic covariance. The evolution of the Sliced-Wasserstein distance between iterates and the target reflects again that convergence is faster for learning rates close to 1, especially for the case $\alpha = 0.1$. The corresponding slowly decreasing learning rate leads to faster convergence than the fixed learning rate $\gamma_k \equiv 1$ ($\alpha = 0$) in all our experiments on discrete samples.



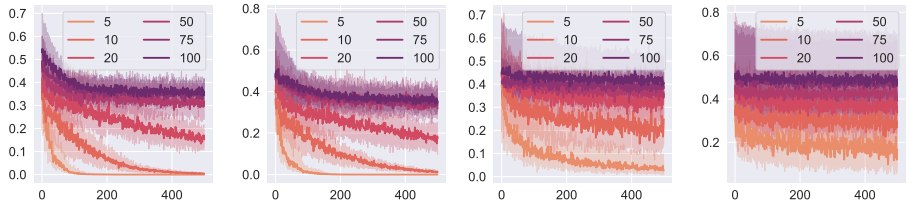
(a) Setting (b) $\alpha = 0$ (c) $\alpha = 0.1$ (d) $\alpha = 0.51$ (e) $\alpha = 0.9$

Figure 6: Evolution of $SW_2^2(\sigma_k, \mu)$ for discrete source and target distributions. The source is sampled from a mixture of Gaussians, and the target is sampled from $\mathcal{N}(0, \mathbf{I}_d)$



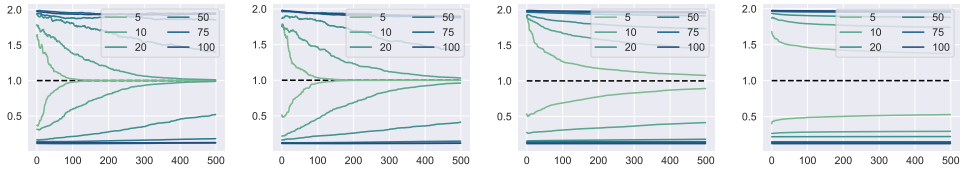
(a) Setting (b) $\alpha = 0$ (c) $\alpha = 0.1$ (d) $\alpha = 0.51$ (e) $\alpha = 0.9$

Figure 7: Evolution of $SW_2^2(\sigma_k, \mu)$ for discrete source and target distributions. The source is sampled from a mixture of Gaussians, and the target is sampled from $\mathcal{N}(0, \Lambda)$



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.51$ (d) $\alpha = 0.9$

Figure 8: Evolution of $SW_2^2(\sigma_k, \mu)$ when $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \mathbf{I}_d)$, with slice-matching maps along a single direction θ_{k+1} instead of an orthonormal basis P_{k+1}



(a) $\alpha = 0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.51$ (d) $\alpha = 0.9$

Figure 9: Minimum and maximum eigenvalues of the estimated covariances Σ_k when $\sigma = \mathcal{N}(0, \Sigma)$ and $\mu = \mathcal{N}(0, \mathbf{I}_d)$, with slice-matching maps along a single direction θ_{k+1} instead of an orthonormal basis P_{k+1}

F.3. A single direction for the slice-matching scheme

Figure 8 and Figure 9 provide alternative experiments when one replaces the orthonormal set of directions P_{k+1} by a single direction θ_{k+1}. We consider continuous Gaussian source and target distributions, so that iterates are explicit. Figure 8 shows the evolution of the Sliced-Wasserstein loss for this alternative algorithm, and Figure 9 shows the evolution of the min/max eigenvalues. Each considers $N = 10$ independent runs, each with a different source covariance. This illustrates how the convergence is worsened for all learning rates and all dimensions d, as compared to our experiments with multiple directions P_{k+1}.