Inexact JKO and proximal-gradient algorithms in the Wasserstein space

Simone Di Marino[*][†] Emanuele Naldi[*][‡] Silvia Villa[*][§]

Abstract

This paper studies the convergence properties of the inexact Jordan-Kinderlehrer-Otto (JKO) scheme and proximal-gradient algorithm in the context of Wasserstein spaces. The JKO scheme, a widely-used method for approximating solutions to gradient flows in Wasserstein spaces, typically assumes exact solutions to iterative minimization problems. However, practical applications often require approximate solutions due to computational limitations. This work focuses on the convergence of the scheme to minimizers for the underlying functional and addresses these challenges by analyzing two types of inexactness: errors in Wasserstein distance and errors in energy functional evaluations. The paper provides rigorous convergence guarantees under controlled error conditions, demonstrating that weak convergence can still be achieved with inexact steps. The analysis is further extended to proximal-gradient algorithms, showing that convergence is preserved under inexact evaluations.

Keywords. JKO scheme; Inexact optimization; Proximal-gradient algorithm; Optimal transport

1 Introduction

The Proximal Point Algorithm (PPA) in the 2-Wasserstein space, also known as Jordan-Kinderlehrer-Otto (JKO) scheme [47] or Minimizing Movement scheme [4, 41], is a well-known variational method for approximating solutions to gradient flows in the Wasserstein space, and as such it is often used in the study of partial differential equations (PDEs) via optimal transport. Given a functional $\mathcal{G}$ defined over the space of probability measures, the JKO scheme approximates the gradient flow of $\mathcal{G}$ in the Wassertein space by iteratively solving a sequence of minimization problems of the form

$$\mu_{n+1} = J_{\tau_n}(\mu_n) := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{G}(\nu) + \frac{1}{2\tau_n} W_2^2(\nu, \mu_n) \right\}, \tag{1.1}$$

where W_2 denotes the 2-Wasserstein distance, $\{\tau_n\}_n$ is a positive sequence of stepsize parameters, and $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability measures with finite second moments.

[*] Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy.
[†] simone.dimarino@unige.it
[‡] emanuele.naldi@edu.unige.it
[§] silvia.villa@unige.it

While the first use of the JKO scheme was to prove existence results for nonlinear PDEs and their convergence properties, the importance of studying JKO schemes extends beyond purely theoretical considerations. In recent years, gradient flows in the space of probability measures have found applications in machine learning [61, 70], neural networks optimization [25, 53, 57], and sampling [10, 24, 35, 69]. As an example, Langevin dynamics describe the evolution of a probability distribution according to a stochastic differential equation that models the behavior of particles in a potential field with added Brownian motion. The Langevin equation can be interpreted as a gradient flow of the Kullback-Leibler (KL) divergence with respect to the Wasserstein-2 metric [69]. The JKO scheme and the proximal-gradient algorithm in Wasserstein spaces [63] provide natural ways to discretize Langevin dynamics by iteratively solving proximal minimization problems related to the KL divergence. Although discretized Wasserstein gradient flows have been proposed in the literature [4, 11, 21, 47, 56, 69], most of them have not been studied as minimization algorithms. In [58] weak convergence results for JKO were established. In [63] estimates for the convergence of a proximal-gradient algorithm were established and later extended to weak convergence in the convex case in [34] only for the Bures–Wasserstein space, that is the (closed and geodesically convex) subset of Gaussians in the Wasserstein space.

Unfortunately, in Wasserstein spaces there are very few cases where the minimization problem (1.1) can be computed in closed form. A notable example is the case where $\mathcal{G}(\mu) = \int g \,\mathrm{d}\mu$ and g is proper, convex, lower semicontinuous and its proximity operator can be computed in a closed form[1], see [11]. Other cases are when the input and the functional are specific, for example when $\mathcal{G}$ is the negative entropy and the input μ is Gaussian, see [69, Example 8].

For this reason, in the Wasserstein setting it is necessary to devise algorithms to compute inexact JKOs, where the exact solution to the minimization problem is replaced by an approximate one. A first option is to regularize problem (1.1), making it possibly easier to solve. One way is to substitute the Wasserstein distance W_2 in (1.1) with its entropic regularized counterpart W_2^ϵ (see [48, 54]). The resulting JKO scheme is analyzed in [20]. However, the main focus of the existing literature is about constructing a sequence close to the Wasserstein Gradient Flow. Other recent strategies rely on regularizing the associated Hamilton-Jacobi equation, motivated by regularized proximal operators in Euclidean space [45, 59], and applying Hopf-Cole type transformations to obtain an effective solution strategy, see [51]. Another regularization strategy can be found in [52], while modern ways to approximate (1.1), involving neural networks, are introduced in [50]. In practice, approximate solutions are unavoidable, since numerical solvers typically achieve a solution only up to a specified tolerance. Optimization schemes that tackle directly the original problem usually make use of the Benamou-Brenier formula [8] to rewrite the 2-Wasserstein distance and solve a saddle point problem. A well-known method to approximate a solution of (1.1) is the so-called "ALG2" introduced in [9], which makes use of an ADMM-type algorithm, but other solvers can be applied to the saddle point problem [23]. In all these cases, the solution to (1.1) is computed only approximately. Despite their practical relevance, the existing convergence analyses for JKO-based algorithms predominantly assume that $\mu_{n+1} = J_{\tau_n}(\mu_n)$, that is, each JKO step is computed exactly. However, as previously discussed, in all existing JKO-based algorithms

[1] See for example the repository at www.proximity-operator.net.

it holds instead

$$\mu_{n+1} \approx J_{\tau_n}(\mu_n).$$

For this reason, understanding the behavior of inexact JKO schemes and quantifying their convergence properties is a crucial step towards broader applicability.

1.1 Inexact JKO

In this work, we address this gap by analyzing two types of errors in the computation of the JKO scheme. The first type involves an error with respect to the Wasserstein distance, where the approximate solution μ_{n+1} satisfies

$$W_2(\mu_{n+1}, J_{\tau_n}(\mu_n)) \leq \epsilon_n, \tag{1.2}$$

with $J_{\tau_n}(\mu_n)$ being the exact JKO step defined in (1.1) and $\{\epsilon_n\}_n$ a nonnegative sequence. We refer to this discrepancy as a *distance-type error.* Our first main result concerns the convergence behavior under this type of approximation.

Theorem 1.1 (Convergence for distance-type error)**.** *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be proper, lower semicontinuous and convex along generalized geodesics with* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ with $\sum_{n=0}^{\infty} \epsilon_n < \infty$ and let $\{\tau_n\}_n \subset \mathbb{R}_{>0}$ with $\sum_{i=0}^{\infty} \tau_i = \infty$. Define $\sigma_n := \sum_{i=0}^{n-1} \tau_i$, for $n \in \mathbb{N}$ and suppose $\sum_{n=1}^{\infty} \frac{\sigma_n}{\tau_n} \epsilon_{n-1}^2 < \infty$. Let $\{\mu_n\}_n$ satisfying* (1.2)*, then*

$$\mathcal{G}(J_{\tau_n}(\mu_n)) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \text{for } n \to \infty,$$

and $W_p(\mu_n, \mu^*) \to 0$ for all $p \in [1, 2)$, where $\mu^* \in \arg\min \mathcal{G}$.

This establishes convergence for a first inexact version of the JKO algorithm, under assumptions on the stepsize and error sequences that are analogous to those used in the analysis within Hilbert spaces. For the sequence $\{J_{\tau_n}(\mu_n)\}_n$, we are able to derive convergence rates in terms of the objective functional $\mathcal{G}$ in relation to the sequence of partial sums of stepsizes $\{\sigma_n\}_n$, mirroring what is known for Hilbert spaces. While this first choice of error seems natural, a second slightly more restrictive choice can be more expressive. In particular, in the next theorem we provide convergence rates for the sequence $\{\mu_n\}_n$ which is the real sequence we have actually access to. Moreover, algorithms that try to solve the minimization problem in (1.1), have as objective to minimize the energy $\mathcal{G}(\cdot) + \frac{1}{2\tau_n} W_2^2(\cdot, \mu_n)$ and can sometimes provide convergence properties in terms of this fuctional. For this reason, the second type of error we consider, is thus measured in terms of the energy of the minimization problem in (1.1) and we refer to it as the *variational-type error*

$$\mathcal{G}(\mu_{n+1}) + \frac{1}{2\tau_n} W_2^2(\mu_{n+1}, \mu_n) \leq \mathcal{G}(J_{\tau_n}(\mu_n)) + \frac{1}{2\tau_n} W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \epsilon_n^2. \tag{1.3}$$

For this error, we will obtain the following result.

Theorem 1.2 (Convergence for variational-type error)**.** *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ proper, lower semicontinuous and convex along generalized geodesics with* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$*

with $\sum_{n=0}^{\infty} \epsilon_n < \infty$*,* $\{\tau_n\}_n \subset \mathbb{R}_{>0}$ *with* $\sum_{i=0}^{\infty} \tau_i = \infty$ *and let* $\sigma_n := \sum_{i=0}^{n-1} \tau_i$*, for* $n \in \mathbb{N}$ *and* $\sum_{n=1}^{\infty} \frac{\sigma_n}{\tau_n} \epsilon_n^2 < \infty$*. Let* $\{\mu_n\}_n$ *satisfying* (1.3)*, then*

$$\mathcal{G}(\mu_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{for } n \to \infty,$$

and $W_p(\mu_n, \mu^*) \to 0$ *for all* $p \in [1, 2)$*, where* $\mu^* \in \arg\min \mathcal{G}$*.*

In contrast to the statement of Theorem 1.1, Theorem 1.2 shows that variational-type errors allow us to establish convergence for the sequence of the values $\mathcal{G}(\mu_n)$ directly, rather than for $\mathcal{G}(J_{\tau_n}(\mu_n))$. This distinction is practically significant, as $J_{\tau_n}(\mu_n)$ is not available in actual computations, whereas μ_n is the output actually produced by the algorithm.

1.2 Inexact proximal-gradient

In the last part of the paper, we focus on the composite problem

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu) = \mathcal{E}_F(\mu) + \mathcal{H}(\mu) \tag{1.4}$$

with $\mathcal{E}_F(\mu) = \int F \, \mathrm{d}\mu$, $F : \mathbb{R}^d \to \mathbb{R}$. The prototypical example of the functional $\mathcal{G}$ is the free energy functional (a key example also in [63]) and it is intimately related to the Kullback-Leibler divergence. The free energy is expressed as

$$\mu \mapsto \int F(x) \, d\mu(x) + \mathrm{Ent}(\mu),$$

where $\mathrm{Ent}(\mu) = \int \log(\mu(x)) d\mu(x)$ if μ is absolutely continuous with respect to the Lebesgue measure and has density μ, and $\mathrm{Ent}(\mu) = +\infty$ otherwise. The entropy functional is proper, lower semicontinuous and convex along generalized geodesics and by definition, its domain satisfies $\mathrm{dom}(\mathrm{Ent}) \subset \mathcal{P}_2^r(\mathcal{X})$, hypothesis that we will assume on $\mathcal{H}$. As highlighted in [63], this example is related to the Langevin dynamic and the Fokker-Planck equation.

The proximal gradient algorithm, originally devised in Hilbert spaces to minimize sums of a smooth and a nonsmooth function similarly to (1.4) has been extended to address problems on the Wasserstein space. This extension was first introduced in [69] and further explored in [63]. The method consists in two alternating steps: a gradient-descent (forward) step for the functional $\mathcal{E}_F$ and a proximal (backward) step for $\mathcal{H}$, and can be written as $\mu_{n+1} = J_{\tau,\mathcal{H}}\left((I - \tau\nabla F)_{\#}(\mu_n)\right)$. Clearly this scheme generalizes the proximal point method (JKO scheme), and the additional assumption we take on the domain of $\mathcal{H}$ is the sole reason for separating the analysis of these two algorithms. However, as we already observed for the JKO scheme, in practice it is generally not feasible to implement an exact proximal-gradient algorithm for this functional (see [69, Section 4.1]) . This limitation then strongly motivates again the introduction of an inexact proximal-gradient scheme.

We consider inexact schemes that perform iterations of the type

$$\mu_{n+1} \approx J_{\tau_n,\mathcal{H}}\left((I - \tau_n \nabla F)_{\#}(\mu_n)\right),$$

with a positive sequence of stepsizes $\{\tau_n\}_n$. In this case, the introduction of a variable stepsize can have practical relevance. In fact, Wibisono notes in [69, Section 2.2.2] that the classical unadjusted Langevin algorithm (ULA) can be interpreted as performing a gradient step for the potential energy and a “flow” step for the entropy functional. Since for small stepsizes the JKO step closely approximates the flow step [4], the whole ULA iteration can actually be interpreted as an inexact step of a proximal-gradient algorithm in Wasserstein spaces. However, it is well-known that the ULA procedure introduces a bias and the algorithm converges to an incorrect distribution. This drawback motivates the analysis of variable (vanishing) stepsizes, as they are sometimes used in practice to “adjust” the ULA scheme and drive convergence towards the correct distribution.

For the convergence analysis of the scheme, we build upon the results of [63], extending them to the convex (and not necessarily strongly convex) setting and to the inexact setting. We establish analogues of Theorem 1.1 and Theorem 1.2 also for the inexact proximal-gradient algorithm, providing convergence guarantees for the resulting sequence $\{\mu_n\}_n$ along with corresponding convergence rates.

For both the algorithms we consider in this work, the results we provide have a long history in Hilbert spaces. In the original works of Martinet [55] and Rockafellar [62], convergence for proximal point algorithms with summable errors were introduced. The impact of errors on convergence has been further analyzed in [5, 28, 44, 64–66]. The extension to Banach spaces and Bregman divergences has also been considered in the works [2, 16, 36] while an analysis of inexact evaluations for proximal-gradient algorithms can be found in [27, 33, 67]. The analysis in Wasserstein spaces, however, is more subtle, as we will discuss throughout the paper. In particular, in Wasserstein spaces it is not possible to rely on the nonexpansivity propriety of the operator J_{τ_n} and thus it is not possible to use directly the correspondent of classical analysis in Hilbert spaces [26].

1.3 Contributions and structure of the paper

The main contributions of this paper can be summarized as follows:

- We propose an inexact JKO framework, where the minimization problems solved at each step allow for controlled approximations in either the Wasserstein distance or energy functional evaluation. We rigorously analyze the convergence of the resulting schemes and provide sufficient conditions for weak convergence. We also discuss rates on the objective functional $\mathcal{G}$.
- We extend the analysis to proximal-gradient algorithms in Wasserstein spaces. Based on the work in [63] we first provide a finer discrete EVI for the proximal-gradient algorithm. With this, we demostrate how the inexact evaluation of proximal steps can still guarantee convergence under suitable assumptions on the error sequence. The results we obtain expand the ones in [63] and [34, Theorem 5.3].
- Both the inexact JKO and proximal-gradient are analyzed with varying stepsizes, which result in a new and interesting analysis, parallel, but with some key differences, to the classical one in Hilbert spaces.

Even if the proximal-gradient algorithm in Wasserstein spaces is more general than the JKO scheme, we decided to keep the contribution separated. The reason lies in the fact that in the analysis of the proximal-gradient algorithm we assume an additional regularity assumption (the same present also in [63]), while we do not need such assumption for the analysis of the JKO algorithm.

The remainder of this paper is organized as follows. In Section 2, we introduce the theoretical background on optimal transport, gradient flows in Wasserstein spaces, and the classical JKO scheme, while also introducing the weak topology we consider in this paper. In this section, we also provide in Theorem 2.13 generalizations of known results to fit our setting. Section 3 introduces the inexact JKO framework, detailing the types of errors considered and presenting the main convergence results. In Section 4, we extend our approach to proximal-gradient algorithms in Wasserstein spaces and establish convergence guarantees for inexact proximal-gradient iterations.

2 Preliminaries

We denote by $\mathcal{M}(\mathbb{R}^d)$ the Banach space of bounded measures defined on $\mathcal{B}(\mathbb{R}^d)$, the Borel σ-algebra of $\mathbb{R}^d$. We define $\mathcal{M}_2(\mathbb{R}^d)$ as the subspace of measures with finite second moments, i.e.,

$$\mathcal{M}_2(\mathbb{R}^d) = \left\{ \mu \in \mathcal{M}(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}|\mu|(x) < +\infty \right\},$$

where $|\mu|$ denotes the total variation of μ. We write $\mathcal{P}(\mathbb{R}^d)$ for the subset of $\mathcal{M}(\mathbb{R}^d)$ of probability measures (i.e., positive measures with mass 1) and we define $\mathcal{P}_2(\mathbb{R}^d) := \mathcal{P}(\mathbb{R}^d) \cap \mathcal{M}_2(\mathbb{R}^d)$.

For every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $L^2(\mu)$ denotes the space of functions $f : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\int \|f\|^2 \,\mathrm{d}\mu < +\infty$. For every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $\|\cdot\|_\mu$ and $\langle \cdot, \cdot \rangle_\mu$ respectively the norm and the inner product of the space $L^2(\mu)$. For any measures μ, ν, we write $\mu \ll \nu$ if μ is absolutely continuous with respect to ν, and we denote $\mathcal{L}^d$ the Lebesgue measure over $\mathbb{R}^d$. The set of absolutely continuous measures with respect to Lebesgue, within $\mathcal{P}_2(\mathbb{R}^d)$, is denoted by $\mathcal{P}_2^r(\mathbb{R}^d) := \{\mu \in \mathcal{P}_2(\mathbb{R}^d), \mu \ll \mathcal{L}^d\}$. Throughout the paper we will also refer to such measures as regular measures.

For every measurable map $T : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ and for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $T_\# \mu \in \mathcal{P}_2(\mathbb{R}^m)$ the pushforward measure of μ by T characterized by the 'transfer lemma', i.e., $\int_{\mathbb{R}^m} \varphi(y) dT_\#\mu(y) = \int_{\mathbb{R}^d} \varphi(T(x)) d\mu(x)$, for any measurable and bounded function φ.

Let $p \in [1, 2]$ and consider the p-Wasserstein distance defined for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$W_p^p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \,\mathrm{d}\gamma(x, y), \tag{2.1}$$

where $\Gamma(\mu, \nu)$ is the set of couplings between μ and ν [68], i.e., the set of nonnegative measures γ over $\mathbb{R}^d \times \mathbb{R}^d$ such that $\pi^1_\# \gamma = \mu$ (respectively $\pi^2_\# \gamma = \nu$) where $\pi^1 : (x, y) \mapsto x$ (respectively $\pi^2 : (x, y) \mapsto y$) is the projection onto the first (respectively second) component. The set $\Gamma(\mu, \nu)$ is called the set of transport plans between μ and ν and the set of plans that minimize (2.1) is the set of optimal transport plans and denoted by $\Gamma_{\mathrm{opt}}(\mu, \nu)$. By Jensen inequality, it is clear that $W_p(\mu, \nu) \leq W_q(\mu, \nu)$ for every $p \leq q$ and $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$.

We recall a foundamental theorem, due to Brenier [15] and Knott-Smith [49], see also [3, Proposition 5.2].

Theorem 2.1 (Brenier, Knott-Smith)**.** *Let $p = 2$, $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

(i) Problem (2.1) *has a unique solution γ. In addition, γ is induced by a transport map, that is, there exists a uniquely determined μ-almost everywhere map $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ such that $\gamma = (I, T_\mu^\nu)_\# \mu$ where $(I, T_\mu^\nu) : x \mapsto (x, T_\mu^\nu(x))$. Moreover $T_\mu^\nu = \nabla\psi$, where $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ is a lower semicontinuous convex function differentiable μ-almost everywhere. The map T_μ^ν is called the optimal transport map from μ to ν.*

(ii) Conversely, if ψ is convex, lower semicontinuous, and differentiable μ-almost everywhere with $|\nabla\psi| \in L^2(\mu)$ and $(\nabla\psi)_\#\mu = \nu$, then $T_\mu^\nu := \nabla\psi$ is the optimal transport map from μ to ν.

(iii) If also $\nu \in \mathcal{P}_2^r(\mathbb{R}^d)$, then $T_\nu^\mu \circ T_\mu^\nu = I$ μ-almost everywhere and $T_\mu^\nu \circ T_\nu^\mu = I$ ν-almost everywhere.

Corollary 2.2. *In the hypothesis of the previous theorem it holds*

$$W_2^2(\mu, \nu) = \inf_{T : T_\#\mu = \nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 d\mu(x).$$

In this paper, as it is commonly the case in the literature, we may refer to the space of probability distributions $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the 2-Wasserstein distance as the Wasserstein space. In the following we define some weak topologies that can be considered on the space $\mathcal{P}_2(\mathbb{R}^d)$.

Definition 2.3 (Narrow topology)**.** *The narrow topology on $\mathcal{P}_2(\mathbb{R}^d)$ is the weak* topology of $(C_b(\mathbb{R}^d))'$, where*

$$C_b(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; f \text{ is continuous and bounded} \right\},$$

endowed with the norm $\|f\|_{C_b(\mathbb{R}^d)} := \sup_{x \in \mathbb{R}^d} |f(x)|$.

This topology is weaker than the strong topology induced by W_p, $p \in [1, 2]$. In particular, we recall that whenever a sequence converges with respect to the distance W_p it converges also narrowly, see [4, Lemma 5.1.7]. We introduce next another topology which will allow us to state more general results throughout the paper. See for example [58, Section 3] and [6, Section 5] for further comments.

Definition 2.4. *Let $C_2^w(\mathbb{R}^d)$ be the space defined by*

$$C_2^w(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; f \text{ is continuous and } \lim_{\|x\| \to \infty} \frac{f(x)}{1 + \|x\|^2} = 0 \right\},$$

endowed with the norm $\|f\|_{C_2^w(\mathbb{R}^d)} := \sup_{x \in \mathbb{R}^d} \frac{|f(x)|}{1 + \|x\|^2}$.

It is known that $\mathcal{M}_2(\mathbb{R}^d)$ can be seen as the dual of such space when endowed with the norm $\|\mu\|_{\mathcal{M}_2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}|\mu|(x)$, see for example [6, Section 5] for a proof and further comments. In this work we consider the weak-$*$ topology in this space restricted to the subset $\mathcal{P}_2(\mathbb{R}^d)$ and we denote it by $\tau_{w,2}$. Whenever a sequence $\{\mu_n\}_n \subset \mathcal{P}_2(\mathbb{R}^d)$ is converging to $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to this topology, we denote it by $\mu_n \overset{w,2}{\rightharpoonup} \mu$. Clearly, this topology is finer than the narrow topology. This means that if a sequence $\{\mu_n\}_n$ is converging to μ in $\mathcal{P}_2(\mathbb{R}^d)$ with respect to the topology $\tau_{w,2}$, then $\mu_n \to \mu$ narrowly. Moreover, the convergence we obtain is "weak" by name, but it implies convergence in p-Wasserstein distance for any $p \in [1, 2)$, see [4, Remark 7.1.11].

Lemma 2.5. *Let $\{\mu_n\}_n$ and $\{\bar{\mu}_n\}_n$ be two bounded sequences in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ such that $W_2(\bar{\mu}_n, \mu_n) \to 0$ and $\bar{\mu}_n \overset{w,2}{\rightharpoonup} \mu^*$, then also $\mu_n \overset{w,2}{\rightharpoonup} \mu^*$.*

Proof. Since both $\{\bar{\mu}_n\}_n$ and $\{\mu_n\}_n$ are bounded sequences in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, there exists $c > 0$ such that

$$\{\bar{\mu}_n\}_n \cup \{\mu_n\}_n \subset K_c := \left\{ \mu \in \mathcal{P}_2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x) \le c \right\}.$$

By [58, Corollary 3.6 (c)] the set K_c is relatively sequentially compact in $\mathcal{P}_2(\mathbb{R}^d)$ endowed with the topology $\tau_{w,2}$. From [4, Lemma 5.1.7] we also have that it is sequentially closed, so that $(K_c, \tau_{w,2})$ is sequentially compact. By hypothesis we have $W_2(\bar{\mu}_n, \mu_n) \to 0$, which implies $W_1(\bar{\mu}_n, \mu_n) \to 0$. Since $\bar{\mu}_n \overset{w,2}{\rightharpoonup} \mu^*$ then $W_1(\bar{\mu}_n, \mu^*) \to 0$ and from the fact that W_1 is induced by a norm, we obtain $W_1(\mu_n, \mu^*) \to 0$. Now, from every subsequence of $\{\mu_n\}_n$ we can extract a further subsequence converging with respect to the topology $\tau_{w,2}$ to some $\mu^{**} \in K_c$, and thus also converging with respect to W_1. However, since (K_c, W_1) is Hausdorff and $W_1(\mu_n, \mu^*) \to 0$, we can conclude that $\mu^{**} = \mu^*$, so that $\mu_n \overset{w,2}{\rightharpoonup} \mu^*$. □

Theorem 2.6 (Opial property,[58, Theorem 5.1])**.** *Let $\{\mu_n\}_n$ such that $\mu_n \overset{w,2}{\rightharpoonup} \mu$ in $\mathcal{P}_2(\mathbb{R}^d)$. Then*

$$W_2^2(\nu, \mu) + \liminf_{k\to\infty} W_2^2(\mu_n, \mu) \le \liminf_{k\to\infty} W_2^2(\mu_n, \nu) \quad \textit{for every } \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

From [58, Corollary 5.3] we have that such property holds under the weaker condition $\mu_n \to \mu$ narrowly in $\mathcal{P}_2(\mathbb{R}^d)$.

Definition 2.7 (Geodesic)**.** *A (minimal, constant speed) geodesic $(\mu_t)_{t\in[0,1]}$ in $\mathcal{P}_2(\mathbb{R}^d)$ connecting two given measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ is a Lipschitz curve satisfying*

$$W_2(\mu_s, \mu_t) = |t - s| W_2(\mu_0, \mu_1) \quad \textit{for every } s, t \in [0, 1]. \tag{2.2}$$

Definition 2.8 (Generalized geodesic)**.** *A generalized geodesic between $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ (with base $\nu \in \mathcal{P}_2(\mathbb{R}^d)$) is a curve $(\mu_t)_{t\in[0,1]}$ in $\mathcal{P}_2(\mathbb{R}^d)$ defined by*

$$\mu_t = (\pi_t^{2\to 3})_{\#}\gamma \quad t \in [0, 1],$$

where $\pi_t^{2\to 3} := (1 - t)\pi^2 + t\pi^3$, $\gamma \in \Gamma(\nu, \mu_0, \mu_1)$, $\pi^{1,2}_{\#}\gamma \in \Gamma_{opt}(\nu, \mu_0)$ and $\pi^{1,3}_{\#}\gamma \in \Gamma_{opt}(\nu, \mu_1)$, with $\pi^{1,2} : (x, y, z) \to (x, y)$ and $\pi^{1,3} : (x, y, z) \to (x, z)$.

Definition 2.9 (Convexity)**.** *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be a proper function. $\mathcal{G}$ is convex along (generalized) geodesics if for every $\mu_0, \mu_1, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a (generalized) geodesic $(\mu_s)_{s\in[0,1]}$ between μ_0 and μ_1 (with base ν), along which*

$$\mathcal{G}(\mu_s) \leq (1-s)\mathcal{G}(\mu_0) + s\mathcal{G}(\mu_1) \quad \text{for every } s \in [0,1]. \tag{2.3}$$

Clearly, convexity along generalized geodesics implies convexity along geodesics. Notice that in [4, Definition 9.2.4], the authors consider as base points only $\nu \in \overline{D(\mathcal{G})}$, this is because they consider as inputs of J_τ only probabilities in $\overline{D(\mathcal{G})}$. For our purposes, however, since the input of J_τ will be an approximation of some element in $D(\mathcal{G})$ and can be strictly outside of the domain, we have to assume this slightly more restrictive definition. The definition we provide is actually common in the literature where some times convexity along generalized geodesics is required for base points $\nu \in \mathcal{P}_2^r(\mathbb{R}^d)$ (see for example [63, Section 2.3.2]). The notion we require coincides with this last one whenever $\mathcal{G}$ is supposed lower semicontinuous, which we will suppose throughout the paper.

Example. *A classical example of a functional that is convex along generalized geodesics is the potential energy*

$$\mu \mapsto \int F(x)\, d\mu(x),$$

where $F : \mathbb{R}^d \to (-\infty, +\infty]$ is proper convex and lower semicontinuous. Another classical example is the interaction energy

$$\mu \to \iint W(x,y) d\mu(x)\, d\mu(y),$$

where $W : \mathbb{R}^d \times \mathbb{R}^d \to (-\infty, +\infty]$ is proper convex and lower semicontinuous. A typical choice is $W(x,y) = V(x-y)$, with $V : \mathbb{R}^d \times \mathbb{R}^d \to (-\infty, +\infty]$ proper convex and lower semicontinuous. A further relevant example we will refer again later is the entropy functional

$$\mu \mapsto Ent(\mu) := \begin{cases} \int \log(\rho(x))\, d\rho(x) & \text{if } \mu \in \mathcal{P}_2^r(\mathbb{R}^d) \text{ and } \mu = \rho\mathcal{L}^d, \\ +\infty & \text{otherwise.} \end{cases}$$

See [4, Section 9.3 and 9.4] for further examples and comments about geodesically convex functionals.

Remark 2.10. *The functional $W_2(\mu^1, \cdot)$ is not convex along geodesics (and therefore it is also not convex along generalized geodesics). However, as noted in [4, Remark 9.2.8], $W_2^2(\mu^1, \cdot)$ is convex along all generalized geodesics with base μ^1. This property is actually essential for the well-posedness of the JKO scheme.*

Theorem 2.11 ([58, Theorem 4.2])**.** *Every lower semicontinuous and geodesically convex functional $\varphi : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is sequentially lower semicontinuous w.r.t. the topology $\tau_{w,2}$ on $\mathcal{P}_2(\mathbb{R}^d)$.*

Definition 2.12 (Subdifferential,[4, Section 10.1.1])**.** *Let $\mathcal{H}$ be a functional defined on $\mathcal{P}_2^r(\mathbb{R}^d)$ and let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. A function $\xi \in L^2(\mu)$ belongs to the (Fréchet) subdifferential of $\mathcal{H}$ at μ iff*

$$\mathcal{H}(\nu) - \mathcal{H}(\mu) \geq \int \langle \xi(x), T_\mu^\nu(x) - x\rangle\, d\mu(x).$$

When is not generating confusion, we refer to a specific element of the subdifferential of $\mathcal{H}$ at μ as $\nabla_W \mathcal{H}(\mu)$.

Let $\tau > 0$, we define the (in general multivalued) operator $J_\tau : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$, by

$$J_\tau(\mu) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{G}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \mu) \right\}. \tag{2.4}$$

Throughout this paper, we will use a specific property (see equation (2.5) below) of the operator J_τ that enables to find what is called a *discrete EVI* for the JKO sequence. Usually, property (2.5) is stated for any initial point $\mu \in \overline{D(\mathcal{G})}$ (see for example [4, Lemma 9.2.7] and [4, Theorem 4.1.2 (i) and (ii)]). However, in our analysis, the input μ will be only an approximation of a previous JKO iterate and, as such, may not lie within the domain of $\mathcal{G}$ (or its closure). For this reason, we need to establish a more general result, for which we provide a proof.

Theorem 2.13. *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ proper, lower semicontinuous and convex along generalized geodesics with* $\operatorname{arg\,min} \mathcal{G} \neq \emptyset$ *and let $\tau > 0$. Then*

(i) For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the minimization problem in (2.4) *has a unique solution*

(ii) For each $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds

$$W_2^2(J_\tau(\mu), \nu) - W_2^2(\mu, \nu) \leq 2\tau \left(\mathcal{G}(\nu) - \mathcal{G}(J_\tau(\mu)) \right) - W_2^2(J_\tau(\mu), \mu) \tag{2.5}$$

(iii) If μ_n is converging to μ in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, then $W_2(J_\tau(\mu_n), J_\tau(\mu))) \to 0$.

Proof. For the first two points we can refer to [4, Lemma 9.2.7] and [4, Theorem 4.1.2 (i) and (ii)]. As we mentioned above, in [4] they consider only $\mu \in \overline{D(\mathcal{G})}$. For our purposes we need the properties to hold for every μ so we provide a brief proof here.

(i) Let us consider a minimizing sequence ν_n for (2.4). Define $\mathcal{G}_\tau(\nu) = \mathcal{G}(\nu) + \frac{1}{2\tau} W_2^2(\mu, \nu)$ and $\mathcal{G}_\tau^* = \inf \mathcal{G}_\tau$ (note that $\mathcal{G}_\tau^* > -\infty$ since $\mathcal{G}_\tau^* \geq \inf \mathcal{G}$). By definition we have $\mathcal{G}_\tau(\nu_n) = \mathcal{G}_\tau^* + \epsilon_n$, where $\epsilon_n \geq 0$ and $\epsilon_n \to 0$. For any $n, m \in \mathbb{N}$ let us consider ν^t a generalized geodesic between ν_n and ν_m with base μ along which $\mathcal{G}$ is convex. By the convexity of $\mathcal{G}$ and the 1-convexity of $W_2^2(\mu, \cdot)$ along this generalized geodesic (see [4, Remark 9.2.8]), for every $t \in (0, 1)$ we obtain

$$\mathcal{G}_\tau(\nu^t) \leq t\mathcal{G}_\tau(\nu_n) + (1 - t)\mathcal{G}_\tau(\nu_m) - \frac{t(1 - t)}{2\tau} W_2^2(\nu_n, \nu_m).$$

Using that $\mathcal{G}_\tau(\nu^{1/2}) \geq \mathcal{G}_\tau^*$ we get $W_2^2(\nu_n, \nu_m) \leq 4\tau(\epsilon_n + \epsilon_m)$, and so ν_n is a Cauchy sequence in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Thus there exists ν_* such that $\nu_n \to \nu_*$ in W_2. By the lower semicontinuity of $\mathcal{G}$ we conclude that $\mathcal{G}_\tau(\nu_*) \leq \liminf_n \mathcal{G}_\tau(\nu_n) = \mathcal{G}_\tau^*$. A similar calculation shows that the minimizer is unique.

(ii) Let us consider a generalized geodesic ν^t between $J_\tau(\mu)$ and ν with base μ. For every $t \in (0, 1)$, we have

$$\mathcal{G}_\tau(\nu^t) \leq t\mathcal{G}_\tau(J_\tau(\mu)) + (1 - t)\mathcal{G}_\tau(\nu) - \frac{t(1 - t)}{2\tau} W_2^2(J_\tau(\mu), \nu);$$

using that $\mathcal{G}_\tau(\nu^t) \geq \mathcal{G}_\tau^* = \mathcal{G}_\tau(J_\tau(\mu))$ we can write

$$(1-t)\mathcal{G}_\tau(J_\tau(\mu)) \leq (1-t)\mathcal{G}_\tau(\nu) - \frac{t(1-t)}{2\tau} W_2^2(J_\tau(\mu), \nu).$$

Dividing by $(1-t)$ and then letting $t \to 1$ we obtain the desired estimate.

(iii) For every μ, we denote by $\mathcal{G}_\tau^*(\mu)$ the value $\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}_\tau(\mu;\nu) := \mathcal{G}(\nu) + \frac{1}{2\tau} W_2^2(\mu,\nu)$, where now we write explicitely the dependence on μ. By [4, Lemma 3.1.2] the functional $\mathcal{G}_\tau^*$ is continuous on $\mathcal{P}_2(\mathbb{R}^d)$. We define now $\nu_n := J_\tau(\mu_n)$ and $\nu := J_\tau(\mu_n)$. Then, we have

$$\limsup_n \mathcal{G}_\tau(\mu;\nu_n) = \limsup_n \mathcal{G}_\tau(\mu_n;\nu_n) = \lim_n \mathcal{G}_\tau^*(\mu_n) = \mathcal{G}_\tau^*(\mu).$$

This implies that for every $\epsilon > 0$ there exists N_ϵ such that

$$\mathcal{G}_\tau(\mu;\nu_n) \leq \mathcal{G}_\tau^*(\mu) + \epsilon = \mathcal{G}_\tau(\mu;\nu) + \epsilon, \tag{2.6}$$

for every $n \geq N_\epsilon$. We notice that $\nu_n \in D(\mathcal{G})$ for all $n \in \mathbb{N}$. By [4, Lemma 9.2.7], there exists a geodesic ν^t between ν_n and ν, with base point μ, such that

$$\mathcal{G}_\tau(\mu;\nu^t) \leq (1-t)\mathcal{G}_\tau(\mu;\nu_n) + t\mathcal{G}_\tau(\mu;\nu) - \frac{t(1-t)}{2\tau} W_2^2(\nu_n,\nu).$$

From (2.6) we obtain

$$\mathcal{G}_\tau(\mu;\nu^t) \leq \mathcal{G}_\tau(\mu;\nu) + (1-t)\epsilon - \frac{t(1-t)}{2\tau} W_2^2(\nu_n,\nu).$$

Using that $\mathcal{G}_\tau(\mu;\nu) = \mathcal{G}_\tau^*(\mu) = \inf \mathcal{G}_\tau(\mu;\cdot)$ we have

$$t(1-t)W_2^2(\nu_n,\nu) \leq 2\tau(1-t)\epsilon.$$

Dividing by $(1-t)$ and letting $t \to 1$, we obtain $W_2^2(\nu_n,\nu) \leq 2\tau\epsilon$ and $W_2(\nu_n,\nu) \to 0$.

□

3 Inexact JKO

As explained in the introduction, we consider here two different choices of error that we allow to be committed from an iteration to another of the JKO scheme. We consider a proper, lower semicontinuous functional $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ which is convex along generalized geodesics and that satisfies $\arg\min \mathcal{G} \neq \emptyset$.

3.1 Distance-type error

We consider in this section a sequence $\{\mu_n\}_n$ generated computing the output $J_{\tau_n}(\mu_n)$ with a certain precision with respect to the Wassertein distance, i.e., a sequence satisfying the following

$$W_2(\mu_{n+1}, J_{\tau_n}(\mu_n)) \leq \epsilon_n \quad \text{for all } n \in \mathbb{N},$$

where $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ and $\{\tau_n\}_n \subset \mathbb{R}_{>0}$. From this, it follows from the triangular inequality that, for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and every $n \in \mathbb{N}$

$$W_2(\mu_{n+1}, \nu) \leq W_2(J_{\tau_n}(\mu_n), \nu) + W_2(\mu_{n+1}, J_{\tau_n}(\mu_n)),$$

so that

$$W_2(\mu_{n+1}, \nu) \leq W_2(J_{\tau_n}(\mu_n), \nu) + \epsilon_n \quad \text{for all } \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

Lemma 3.1. *Suppose that $\sum_n \epsilon_n < +\infty$. Then for every $\nu \in \arg\min \mathcal{G}$ the sequences $\{W_2(\mu_n, \nu)\}_n$ and $\{W_2(J_{\tau_n}(\mu_n), \nu)\}_n$ converge and there exists a contant $C > 0$ such that*

$$W_2^2(\mu_{n+1}, \nu) \leq W_2^2(J_{\tau_n}(\mu_n), \nu) + C\epsilon_n. \tag{3.1}$$

Proof. For every $\nu \in \arg\min \mathcal{G}$, we have from (2.5)

$$W_2(J_{\tau_n}(\mu_n), \nu) \leq W_2(\mu_n, \nu),$$

so that

$$W_2(\mu_{n+1}, \nu) \leq W_2(\mu_n, \nu) + \epsilon_n,$$

and the sequence $\{W_2(\mu_n, \nu)\}_n$ converges. On the other hand, we also have

$$W_2(J_{\tau_n}(\mu_{n+1}), \nu) \leq W_2(\mu_{n+1}, \nu) \leq W_2(J(\mu_n), \nu) + \epsilon_n,$$

which shows that $\{W_2(J_{\tau_n}(\mu_n), \nu)\}_n$ converges too. It also follows that there exists a $c > 0$ such that $W_2(\mu_n, \nu) \leq c$ for all $n \in \mathbb{N}$ (and thus also $W_2(J_{\tau_n}(\mu_n), \nu) \leq c$ for all $n \in \mathbb{N}$). Thus, we obtain

$$\begin{aligned} W_2^2(\mu_{n+1}, \nu) &= W_2(\mu_{n+1}, \nu) W_2(\mu_{n+1}, \nu) \\ &\leq (W_2(J_{\tau_n}(\mu_n), \nu) + \epsilon_n)(W_2(J_{\tau_n}(\mu_n), \nu) + \epsilon_n) \\ &\leq W_2^2(J_{\tau_n}(\mu_n), \nu) + 2c\epsilon_n + \epsilon_n^2, \end{aligned}$$

and since ϵ_n is bounded, this concludes the proof. □

The previous lemma implies quasi Fejér monotonicity with respect to $\arg\min \mathcal{G}$ (see [7, Definition 5.32] for the definition in Hilbert spaces). Specifically, it follows that for all $\nu \in \arg\min \mathcal{G}$ there exists a constant $C > 0$ such that

$$W_2^2(\mu_{n+1}, \nu) \leq W_2^2(\mu_n, \nu) + C\epsilon_n. \tag{3.2}$$

From now on, we will use the notation $\tilde{\epsilon}_n := C\epsilon_n$, for all $n \in \mathbb{N}$.

Lemma 3.2. *It holds that $W_2(J_{\tau_n}(\mu_n), \mu_n) \to 0$ as $n \to +\infty$.*

Proof. Combining (3.1) with the discrete EVI inequality (2.5), we obtain

$$W_2^2(\mu_{n+1}, \nu) - W_2^2(\mu_n, \nu) \leq 2\tau_n \left(\mathcal{G}(\nu) - \mathcal{G}(J_{\tau_n}(\mu_n))\right) - W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \tilde{\epsilon}_n, \tag{3.3}$$

for all $\nu \in \arg\min \mathcal{G}$. From this, we derive

$$W_2^2(\mu_{n+1}, \nu) \leq W_2^2(\mu_n, \nu) - W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \tilde{\epsilon}_n.$$

Since $\sum_n \tilde{\epsilon}_n = C \cdot \sum_n \epsilon_n < +\infty$, summing up both sides gives

$$\sum_n W_2^2(J_{\tau_n}(\mu_n), \mu_n) < +\infty. \tag{3.4}$$

This implies in particular that $W_2(J_{\tau_n}(\mu_n), \mu_n) \to 0$. □

Theorem 3.3. *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be proper, lower semicontinuous, and convex along generalized geodesics, with* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ with $\sum_{n=0}^{\infty} \epsilon_n < \infty$ and let $\{\tau_n\}_n \subset \mathbb{R}_{>0}$ with $\sum_{i=0}^{\infty} \tau_i = \infty$. Define $\sigma_n := \sum_{i=0}^{n-1} \tau_i$, for $n \in \mathbb{N}$. Let $\{\mu_n\}_n$ be a sequence satisfying*

$$W_2(\mu_{n+1}, J_{\tau_n}(\mu_n)) \leq \epsilon_n, \quad \textit{for all } n \in \mathbb{N}.$$

1. *It holds the rate*

$$\mathcal{G}(\bar{\beta}_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{as } n \to \infty, \tag{3.5}$$

where $\bar{\beta}_n := J_{\tau_{j_n}}(\mu_{j_n})$ with $j_n = \arg\min_{i=0,\dots,n-1}\{\mathcal{G}(J_{\tau_i}(\mu_i))\}$, defines the sequence of the best iterates.

2. *If $\sum_{n=1}^{\infty} \frac{\sigma_n}{\tau_n} \epsilon_{n-1}^2 < \infty$, then*

$$\mathcal{G}(J_{\tau_n}(\mu_n)) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{as } n \to \infty,$$

and $\{\mu_n\}_n$ converges with respect to the topology $\tau_{w,2}$ to some $\mu^* \in \arg\min \mathcal{G}$. In particular we have $W_p(\mu_k, \mu^*) \to 0$ for all $p \in [1, 2)$.

Proof. We first prove that every weak cluster point of $\{J_{\tau_n}(\mu_n)\}_n$ is a minimizer of $\mathcal{G}$. We recall equation (3.3) from which we know it holds

$$W_2^2(\mu_{n+1}, \nu) - W_2^2(\mu_n, \nu) \leq 2\tau_n(\mathcal{G}(\nu) - \mathcal{G}(J_{\tau_n}(\mu_n))) - W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \tilde{\epsilon}_n, \tag{3.6}$$

for all $\nu \in \arg\min \mathcal{G}$ and $n \in \mathbb{N}$. Summing up from 0 to $N-1$, we obtain

$$2\sum_{n=0}^{N-1} \tau_n \mathcal{G}(J_{\tau_n}(\mu_n)) + W_2^2(\mu_N, \nu) \leq 2\sigma_N \mathcal{G}(\nu) + W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} (\tilde{\epsilon}_n - W_2^2(J_{\tau_n}(\mu_n), \mu_n)). \tag{3.7}$$

Since $\bar{\beta}_N$ is the best iterate, we have

$$\mathcal{G}(\bar{\beta}_N) \leq \frac{1}{\sigma_N} \sum_{n=0}^{N-1} \tau_n \mathcal{G}(\mu_{n+1}), \tag{3.8}$$

and combining this with equation (3.7), we obtain

$$\mathcal{G}(\bar{\beta}_N) - \inf \mathcal{G} \leq \frac{C}{\sigma_N}. \tag{3.9}$$

The first part of the theorem is established. On the other hand, for all $n \geq 1$, we have

$$\begin{aligned} \mathcal{G}(J_{\tau_n}(\mu_n)) + \frac{1}{2\tau_n} W_2^2(J_{\tau_n}(\mu_n), \mu_n) &\leq \mathcal{G}(J_{\tau_{n-1}}(\mu_{n-1})) + \frac{1}{2\tau_n} W_2^2(J_{\tau_{n-1}}(\mu_{n-1}), \mu_n) \\ &\leq \mathcal{G}(J_{\tau_{n-1}}(\mu_{n-1})) + \frac{1}{2\tau_n} \epsilon_{n-1}^2, \end{aligned} \tag{3.10}$$

and thus, multiplying by σ_n

$$(\sigma_{n+1} - \tau_n)\mathcal{G}(J_{\tau_n}(\mu_n)) - \sigma_n \mathcal{G}(J_{\tau_{n-1}}(\mu_{n-1})) \leq \frac{\sigma_n}{2\tau_n} \epsilon_{n-1}^2. \tag{3.11}$$

Summing from 1 to $N-1$ and recalling that $\sigma_1 = \tau_0$, we obtain

$$\sigma_N \mathcal{G}(J_{\tau_{N-1}}(\mu_{N-1})) - \sum_{n=1}^{N-1} \frac{\sigma_n}{2\tau_n}\epsilon_{n-1}^2 \leq \sum_{n=0}^{N-1} \tau_n \mathcal{G}(J_{\tau_n}(\mu_n)). \tag{3.12}$$

Combining (3.7) and (3.12), we arrive at

$$2\sigma_N \mathcal{G}(J_{\tau_{N-1}}(\mu_{N-1})) - 2\sigma_N \mathcal{G}(\nu) \leq W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} \tilde{\epsilon}_n + \sum_{n=1}^{N-1} \frac{\sigma_n}{\tau_n}\epsilon_{n-1}^2. \tag{3.13}$$

Since by hypothesis $\tilde{\epsilon}_n$ and $\frac{\sigma_n}{\tau_n}\epsilon_{n-1}^2$ are summable, we get

$$\mathcal{G}(J_{\tau_{N-1}}(\mu_{N-1})) - \inf \mathcal{G} \leq \frac{C}{\sigma_N}, \tag{3.14}$$

for some constant $C > 0$. To conclude we use the Opial property (Theorem 2.6) to prove convergence of the whole sequence. First we notice that $\{J_{\tau_n}(\mu_n)\}_n$ is bounded in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ and thus

$$\sup_{n\in\mathbb{N}} \int \|x\|^2 \, dJ_{\tau_n}(\mu_n)(x) = \sup_{n\in\mathbb{N}} W_2^2(J_{\tau_n}(\mu_n), \delta_0) < +\infty.$$

Using for example [58, Corollary 3.6 (c)] we conclude that $\{J_{\tau_n}(\mu_n)\}_n$ has at least a cluster point with respect to the topology $\tau_{w,2}$ in $\mathcal{P}_2(\mathbb{R}^d)$. Consider a subsequence $\{J_{\tau_{n_i}}(\mu_{n_i})\}_i$ of $\{J_{\tau_n}(\mu_n)\}_n$ and a cluster point $\mu \in \mathcal{P}(X)$ such that $J_{\tau_{n_i}}(\mu_{n_i}) \overset{w,2}{\rightharpoonup} \mu$, then, by lower semicontinuity of $\mathcal{G}$ (also with respect to the topology $\tau_{w,2}$ we consider on $\mathcal{P}^2(\mathbb{R}^d)$, see Theorem 2.11), we obtain

$$\mathcal{G}(\mu) \leq \liminf_i \mathcal{G}(J_{\tau_{n_i}}(\mu_{n_i})) \leq \limsup_i \mathcal{G}(J_{\tau_{n_i}}(\mu_{n_i})) \leq \inf \mathcal{G},$$

which means that $\mu \in \arg\min \mathcal{G}$. This proves that every cluster point of the sequence is a minimizer of $\mathcal{G}$.

Now, in order to prove convergence of the whole sequence, we suppose there exist two subequences $\{J_{\tau_{n_i}}(\mu_{n_i})\}_i$ and $\{J_{\tau_{m_i}}(\mu_{m_i})\}_i$ of $\{J_{\tau_n}(\mu_n)\}_n$ with $J_{\tau_{n_i}}(\mu_{n_i}) \overset{w,2}{\rightharpoonup} \nu^* \in \mathcal{P}_2(\mathbb{R}^d)$ and $J_{\tau_{m_i}}(\mu_{m_i}) \overset{w,2}{\rightharpoonup} \nu^{**} \in \mathcal{P}_2(\mathbb{R}^d)$. We define $\ell(\nu) := \lim_n W_2(J_{\tau_n}(\mu_n), \nu)$ for all $\nu \in \arg\min \mathcal{G}$, which always exists by Lemma 3.1. Since we have proven $\nu^*, \nu^{**} \in \arg\min \mathcal{G}$, we can use the Opial property and obtain whenever $\nu^* \neq \nu^{**}$

$$\ell(\nu^*) = \liminf_i W_2(J_{\tau_{n_i}}(\mu_{n_i}), \nu^*) < \liminf_i W_2^2(J_{\tau_{n_i}}(\mu_{n_i}), \nu^{**}) = \ell(\nu^{**})$$
$$\ell(\nu^{**}) = \liminf_i W_2(J_{\tau_{m_i}}(\mu_{m_i}), \nu^{**}) < \liminf_i W_2^2(J_{\tau_{m_i}}(\mu_{m_i}), \nu^*) = \ell(\nu^*)$$

so that it must be $\nu^* = \nu^{**}$. We have proved that there exists $\mu^* \in \arg\min \mathcal{G}$ such that $J_{\tau_n}(\mu_n) \overset{w,2}{\rightharpoonup} \mu^*$. Since both $\{J_{\tau_n}(\mu_n)\}_n$ and $\{\mu_n\}_n$ are bounded sequences and by Lemma 3.2 we have $W_2(J_{\tau_n}(\mu_n), \mu_n) \to 0$, we can conclude that $\mu_n \overset{w,2}{\rightharpoonup} \mu^*$ using Lemma 2.5. □

Remark 3.4. *Notice that imposing that $\mathcal{G}$ is $\lambda_{\mathcal{G}}$-convex along generalized geodesics [4, Definition 9.2.4], with $\lambda_{\mathcal{G}} > 0$, an analogue of [4, Theorem 4.1.2 (ii)] combined with* (3.6) *yields the inequality*

$$(1 + \lambda_{\mathcal{G}}\tau_n) W_2^2(\mu_{n+1}, \nu) \leq W_2^2(\mu_n, \nu) + \tilde{\epsilon}_n \qquad \textit{for all } \nu \in \arg\min \mathcal{G}.$$

From this, strong convergence in W_2 can be obtained. However, due to the error ϵ_n incurred at each iteration, the convergence rate cannot be improved without further assumptions on $\{\epsilon_n\}_n$. This consideration motivates our decision not to treat the strongly convex case separately.

Remark 3.5. *The condition $\sum_{n=0}^{\infty} \frac{\sigma_n}{\tau_n} \epsilon_{n-1}^2 < \infty$ is not overly restrictive and it is somehow what we expect. This condition is satisfied in particular in the following cases:*

- $\epsilon_n = 0$ *for all* $n \in \mathbb{N}$
- $\{\epsilon_n\}_n$ *is nonincreasing and* $0 < \inf_n \tau_n \leq \sup_n \tau_n < M$ *for some* $M > 0$.
- $\{\epsilon_n\}_n$ *and* $\{\tau_n\}_n$ *nonincreasing and* $\sigma_n \epsilon_{n-1}/\tau_n$ *is bounded, for example with* $\epsilon_n = \frac{1}{n^{1+\delta}}$ *and* $\tau_n = \frac{1}{n}$, *for all* $n \in \mathbb{N}$, *where* $\delta > 0$.
- $\{\epsilon_n\}_n$ *is nonincreasing and* $\{\tau_n\}_n$ *nondecreasing.*

Remark 3.6 (Ergodic convergence)**.** *Supposing that the elements of the sequence $\{J_{\tau_n}(\mu_n)\}_n$ belong to $\mathcal{P}_2^r(\mathbb{R}^d)$ we can apply [1, Proposition 7.6] which states that if a functional is convex along generalized geodesics then it is convex along barycenters (for regular measures). In particular, this implies that* (3.8) *holds for $\bar{\beta}_n := Bar\left(J_{\tau_i}(\mu_i), \frac{\tau_i}{\sigma_n}\right)_{i=0,\dots,n-1}$, the Wasserstein barycenter of the first n elements of the sequence $\{J_{\tau_i}(\mu_i)\}_i$ with parameters $\{\frac{\tau_i}{\sigma_n}\}_0^{n-1}$. As a result,* (3.5) *holds for the barycenter sequence $\{\bar{\beta}_n\}_n$, demonstrating that ergodic-type convergence rates can be achieved without requiring the additional assumptions in point 2 of Theorem 3.3. We expect this result to hold under more general assumptions, for example for functionals for which regular measures are dense in energy. In such cases, one could exploit the stability of Wasserstein barycenters, see [43, Theorem 3] and [19]. We do not pursue this direction since Wasserstein barycenters are not as practical to compute as their Hilbertian counterpart, thereby limiting their practical relevance.*

Remark 3.7. *When $\mathcal{G}$ is $L_{\mathcal{G}}$-Lipschitz continuous in a ball that contains $\{\mu_n\}_n$ and $\{J_{\tau_n}(\mu_n)\}_n$, we can derive a rate on $\mathcal{G}(\mu_n)$. In fact, we have $\mathcal{G}(\mu_{n+1}) \leq \mathcal{G}(J_{\tau_n}(\mu_n)) + L_{\mathcal{G}}\epsilon_n$, and if $\epsilon_n = O\left(\frac{1}{\sigma_n}\right)$, for $n \to \infty$, then*

$$\mathcal{G}(\mu_{n+1}) - \inf \mathcal{G} \leq \mathcal{G}(J_{\tau_n}(\mu_n)) - \inf \mathcal{G} + L_{\mathcal{G}}\epsilon_n = O\left(\frac{1}{\sigma_n}\right), \quad \text{for } n \to \infty.$$

Notice, however, that in Wasserstein spaces, the connection between convexity and local Lipschitz continuity is not obvious. As an example, let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ be proper, lower semicontinuous and convex along geodesics. If $\mathcal{G}$ is approximable by discrete measures and the domain of $\mathcal{G}$ is totally convex, then $\mathcal{G}$ is locally Lipschitz, see [22, Remark 9]. However, several functionals are not locally Lipschitz continuous. In particular, if the domain of $\mathcal{G}$ is included in the set of regular measures $\mathcal{P}_2^r(\mathbb{R}^d)$, then $\mathcal{G}$ cannot be locally Lipschitz continuous since every regular measure can be approximated in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ by discrete measures. A notable example where locally Lipschitz continuity fails is given by the negative entropy.

Nonexpansivity of the proximal map In classical analysis in Hilbert spaces, the weak convergence of the sequence $\{\mu_n\}_n$ is established in a more direct way. However, such an analysis relies on the nonexpansivity of the proximal map, a property that is not known to hold in Wasserstein spaces. Indeed, this remains an open problem. Even the projection onto convex sets can fail to be convex. As a first example, consider the set of two-atomic measures

$$S = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \mid \#\operatorname{supp}(\mu) = 2\}.$$

This set is closed and geodesically convex, but the projection onto this set fails to be Lipschitz, as also shown in Figure 1.

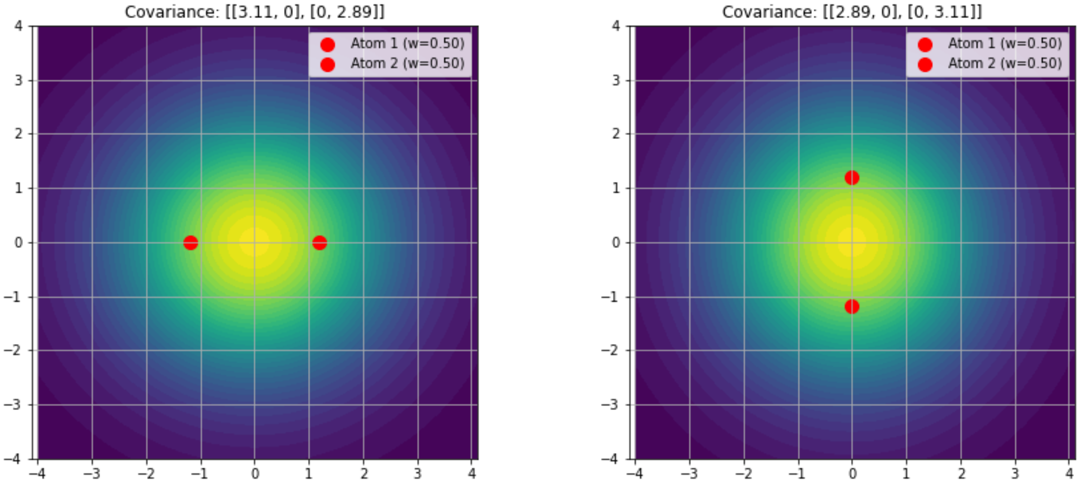

Figure 1: Projection of two nearby 2D Gaussians onto the set S. Despite having really similar distributions, the projections differ significantly, showing that the projection map is not Lipschitz.

Even considering sets which are convex along generalized geodesics, the Lipschitzianity could fail or remain open, as commented in [32]. Consider the set

$$S' := \left\{\rho \in L_1^+(\mathbb{R}^d) \mid \int \rho(x)\,dx = 1,\ \rho \leq 1\right\} \cap \mathcal{P}_2(\mathbb{R}^d)$$

In [32, Corollary 5.3] it is proven that the projection onto S' is locally $\frac{1}{2}$-Hölder, but in [32, Remark 5.1] the authors say that Lipschitzianity is still an open problem.
For positive results, some insights are provided in [18], where a slight modification of the metric is considered and there are specific cases where the nonexpansivity of J_τ is guaranteed. A known positive example of the nonexpansivity of the proximal map, is when the functional $\mathcal{G}$ is totally convex, i.e., convex along any coupling, see [22, Definition 2.7]. In this case, it is possible to introduce a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, consider the function

$$g : L^2(\Omega, \mathbb{P}; \mathbb{R}^d) \to (-\infty, +\infty] : X \to \mathcal{G}(X_\# \mathbb{P}),$$

and prove that g is proper, convex and lower semicontinuous. By the last part of [22, Proposition 5.2 (1)], whenever $\mu = X_{\#}\mathbb{P}$, then $J_\tau(\mu) = (\operatorname{prox}^{L^2}_{\tau g} \circ X)_{\#}\mathbb{P}$. Since g is proper, convex and lower semicontinuous, then $\operatorname{prox}^{L^2}_{\tau g} : L^2(\Omega, \mathbb{P}; \mathbb{R}^d) \to L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ is nonexpansive [REF?]. Thus, given $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, for every X, X' such that $X_{\#}\mathbb{P} = \mu$, $X'_{\#}\mathbb{P} = \mu'$, we have

$$W_2^2(J_\tau(\mu), J_\tau(\mu')) \leq \| \operatorname{prox}^{L^2}_{\tau g}(X) - \operatorname{prox}^{L^2}_{\tau g}(X')\|^2_{L^2} \leq \|X - X'\|^2_{L^2}.$$

So that

$$W_2(J_\tau(\mu), J_\tau(\mu')) \leq \inf_{\substack{X \sim \mu \\ X' \sim \mu'}} \|X - X'\|^2_{L^2} = W_2(\mu, \mu').$$

3.2 Variational-type error

Considering the second error choice described in the introduction, we can prove the following result.

Theorem 3.8. *Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ proper, lower semicontinuous and convex along generalized geodesics and suppose $\arg\min \mathcal{G} \neq \emptyset$. Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ with $\sum_{n=0}^\infty \epsilon_n < \infty$ and let $\{\tau_n\}_n \subset \mathbb{R}_{>0}$ with $\sum_{i=0}^\infty \tau_i = \infty$. Define $\sigma_n := \sum_{i=0}^{n-1} \tau_i$, for $n \in \mathbb{N}$. Consider $\{\mu_n\}_n$ such that for all $n \in \mathbb{N}$ it holds*

$$\mathcal{G}(\mu_{n+1}) + \frac{1}{2\tau_n} W_2^2(\mu_{n+1}, \mu_n) \leq \mathcal{G}(J_{\tau_n}(\mu_n)) + \frac{1}{2\tau_n} W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \frac{\epsilon_n^2}{2\tau_n}. \tag{3.15}$$

1. *It holds the rate*
$$\mathcal{G}(\beta_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \text{as } n \to \infty,$$
where $\beta_n := \mu_{j_n}$ with $j_n = \arg\min_{i=0,\dots,n-1}\{\mathcal{G}(\mu_i)\}$, defines the sequence of the best iterate.

2. *If $\sum_{n=0}^\infty \frac{\sigma_n}{\tau_n}\epsilon_n^2 < \infty$, then*
$$\mathcal{G}(\mu_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \text{as } n \to \infty,$$
and $\{\mu_n\}_n$ converges with respect to the topology $\tau_{w,2}$ to some $\mu^* \in \arg\min \mathcal{G}$. In particular we have $W_p(\mu_k, \mu^*) \to 0$ for all $p \in [1, 2)$.

Proof. Set $\mathcal{G}_{\tau_n}(\mu) := \mathcal{G}(\mu) + \frac{1}{2\tau_n} W_2^2(\mu, \mu_n)$. The assumption (3.15) then reads

$$\mathcal{G}_{\tau_n}(\mu_{n+1}) \leq \mathcal{G}_{\tau_n}(J_{\tau_n}(\mu_n)) + \frac{\epsilon_n^2}{2\tau_n}. \tag{3.16}$$

By [4, Lemma 9.2.7], there exists a generalized geodesic ν^t between μ_{n+1} and $J_{\tau_n}(\mu_n)$, with base point μ_n, such that

$$\mathcal{G}_{\tau_n}(\nu^t) \leq (1-t)\mathcal{G}_{\tau_n}(\mu_{n+1}) + t\mathcal{G}_{\tau_n}(J_{\tau_n}(\mu_n)) - \frac{t(1-t)}{2\tau_n} W_2^2(\mu_{n+1}, J_{\tau_n}(\mu_n)).$$

From (3.16) we derive

$$\mathcal{G}_{\tau_n}(\nu^t) \leq \mathcal{G}_{\tau_n}(J_{\tau_n}(\mu_n)) + (1-t)\frac{2\epsilon_n^2}{\tau_n} - \frac{t(1-t)}{2\tau_n} W_2^2(\mu_{n+1}, J_{\tau_n}(\mu_n))$$

and

$$W_2^2(\mu_{n+1}, J_{\tau_n}(\mu_n)) \leq 2\tau_n \frac{\mathcal{G}_{\tau_n}(J_{\tau_n}(\mu_n)) - \mathcal{G}_{\tau_n}(\nu^t) + (1-t)\frac{\epsilon_n^2}{2\tau_n}}{t(1-t)}.$$

Using that $\mathcal{G}_{\tau_n}(J_{\tau_n}(\mu_n)) = \inf_\mu \mathcal{G}_{\tau_n}(\mu)$ and letting $t \to 1$, we obtain

$$W_2^2(\mu_{n+1}, J_{\tau_n}(\mu_n)) \leq \epsilon_n^2 \quad \text{and} \quad W_2(\mu_{n+1}, J_\tau(\mu_n)) \leq \epsilon_n.$$

We can then apply Theorem 3.3 to obtain convergence of the sequence $\{\mu_n\}_n$ to a minimizer of $\mathcal{G}$. In order to derive the rate on $\mathcal{G}(\mu_n)$, we proceed as follows. We first recall from (3.7) that

$$2\sum_{n=0}^{N-1} \tau_n \mathcal{G}(J_{\tau_n}(\mu_n)) + W_2^2(\mu_N, \nu) \leq 2\sigma_N \mathcal{G}(\nu) + W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} (\tilde{\epsilon}_n - W_2^2(J_{\tau_n}(\mu_n), \mu_n)). \tag{3.17}$$

and thus, assumption (3.15) implies

$$2\sum_{n=0}^{N-1} \tau_n \mathcal{G}(\mu_{n+1}) + W_2^2(\mu_N, \nu) \leq 2\sigma_N \mathcal{G}(\nu) + W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} \tilde{\epsilon}_n + \sum_{n=0}^{N-1} \epsilon_n^2, \tag{3.18}$$

and point 1 follows from the same reasoning done in the proof of Theorem 3.3. On the other hand, (3.15) yields

$$\mathcal{G}(\mu_{n+1}) \leq \mathcal{G}(J_{\tau_n}(\mu_n)) + \frac{1}{2\tau_n} W_2^2(J_{\tau_n}(\mu_n), \mu_n) + \frac{\epsilon_n^2}{2\tau_n} \leq \mathcal{G}(\mu_n) + \frac{\epsilon_n^2}{2\tau_n}.$$

Reasoning as in the proof of Theorem 3.3, we multiply by σ_n and obtain

$$(\sigma_{n+1} - \tau_n)\mathcal{G}(\mu_{n+1}) - \sigma_n \mathcal{G}(\mu_n) \leq \frac{\sigma_n}{2\tau_n} \epsilon_n^2. \tag{3.19}$$

Summing from 0 to $N-1$ and recalling that $\sigma_0 = 0$, we derive

$$\sigma_N \mathcal{G}(\mu_N) - \sum_{n=0}^{N-1} \frac{\sigma_n}{2\tau_n} \epsilon_n^2 \leq \sum_{n=0}^{N-1} \tau_n \mathcal{G}(\mu_{n+1}). \tag{3.20}$$

Combining, we obtain

$$2\sigma_N \mathcal{G}(\mu_N) \leq 2\sigma_N \mathcal{G}(\nu) + W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} \tilde{\epsilon}_n + \sum_{n=0}^{N-1} \epsilon_n^2 + \sum_{n=0}^{N-1} \frac{\sigma_n}{\tau_n} \epsilon_n^2. \tag{3.21}$$

Using the summability conditions, we end up with

$$\mathcal{G}(\mu_N) - \inf \mathcal{G} \leq \frac{C}{\sigma_N}. \tag{3.22}$$

□

Obtaining error bounds approximating the solution Optimization schemes that tackle problem (1.1) usually make use of the Benamou-Brenier formula [8] to rewrite the 2-Wasserstein distance, and solve a saddle point problem. A well-known method to approximate a solution of (1.1) is the so-called "ALG2" introduced in [9]. The algorithm makes use of an augmented Lagrangian method called alternating direction method of multipliers (ADMM). The ADMM method has a long history and presents solid guarantees of convergence [38–40, 42, 46, 60]. For example, in the overview of Boyd [12] the convergence of the "dual" variable is stated in [12, Section 3.2.1]. Using the notations of [9], the "dual" variable correspond to $\sigma = (\mu, m, \mu_1)$. In particular, the convergence of σ_n to the optimal dual variable σ^* implies the convergence of $\mu_{1,n}$ to the solution μ_1^* to problem (1.1) (or [9, Problem (1.4)]). It is clear that both F and G in [9] are not strongly convex, independently on the choice of the initial functional to minimize (in our notations $\mathcal{G}$). Thus, it's not obvious how to guarantee linear convergence for the "dual" variable σ. Energy convergence is also stated in [12, Section 3.2.1], but it is not clear how to find a condition of the kind (3.15). It is possible that results about convergence in energy of the closely related Douglas-Rachford method or its generalization have to be used [13, 14, 30, 31, 37]. This can be the subject of further investigation.

Obtaining error bounds modifying the problem As discussed in the introduction, another common practice is to perturb the original problem (1.1) to find in an easier way approximated solutions. In [20] a perturbation of the W_2-distance coming from entropic optimal transport [29] is used instead of W_2 in (1.1). However, the authors do not quantify the W_2-distance between the output of the entropic proximal step and the output of the classic proximal step, so it is not clear if it is possible to achieve one of the previous conditions, even sending the regularization parameter to zero. However, in these cases, it is probably better to try and prove some sort of convergence result for the scheme itself, without relying on how close it is to a classical JKO scheme. We do not expand on this in the current work. Other approximations of the JKO step can be found in [50–52]. Also here the authors do not provide any estimate about how far the approximated solutions are from the true solution and investigate the quality of the approximations only empirically.

4 Inexact proximal-gradient algorithm

As anticipated in the introduction, in this section we focus on the problem

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{G}(\mu) = \mathcal{E}_F(\mu) + \mathcal{H}(\mu) \tag{4.1}$$

with $\mathcal{E}_F(\mu) = \int F \, \mathrm{d}\mu$. Our analysis is grounded on the following assumptions.

(A1) $F\colon \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable with L-Lipschitz continuous gradient and λ-strongly convex (with $\lambda = 0$ permitted)

(A2) $\mathcal{H}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ is proper, lower semicontinuous and convex along generalized geodesics

(A3) $\mathrm{dom}(\mathcal{H}) \subset \mathcal{P}_2^r(\mathcal{X})$.

Notice that, as an example, the task of minimizing the free energy functional defined by

$$\mu \mapsto \int F(x)\, d\mu(x) + \mathrm{Ent}(\mu),$$

can be cast as (4.1) with $\mathcal{H} = \mathrm{Ent}$, which satisfies assumption (A2)-(A3) since it is proper, lower semicontinuous, convex along generalized geodesics and by definition its domain satisfies $\mathrm{dom}(\mathrm{Ent}) \subset \mathcal{P}_2^r(\mathcal{X})$.

The proximal-gradient algorithm has been studied in [63]. At each iteration makes use of gradient-descent-type step for the functional $\mathcal{E}_F$ and of a proximal step for $\mathcal{H}$. The resulting scheme is a more general algorithm than the proximal point method (JKO scheme), and the additional assumption (A3) is the reason why we keep separated the analysis of the two algorithms. In this section, we describe an inexact version of the method. We define the operator

$$S_\tau := J_{\tau,\mathcal{H}} \circ (I - \tau\nabla F)_\#$$

and we start by considering an inexact scheme satisfying

$$W_2(\mu_{n+1}, S_{\tau_n}(\mu_n)) \leq \epsilon_n, \quad \text{for all } n \in N, \tag{4.2}$$

with a sequence of positive stepsizes $\{\tau_n\}_n$.

First, analyzing [63, Proposition 8] and its proof, it is possible to show that, whenever $\tau < 1/L$, for all $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds

$$W_2^2(S_\tau(\mu), \nu) \leq (1 - \tau\lambda)W_2^2(\mu, \nu) - 2\tau(\mathcal{G}(S_\tau(\mu)) - \mathcal{G}(\nu)). \tag{4.3}$$

It is actually possible to prove a more refined result, which will be crucial.

Lemma 4.1. *Let $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and F satisfying (A1)-(A3). Let $S_\tau := J_{\tau,\mathcal{H}} \circ (I - \tau\nabla F)_\#$ with $\tau < 1/L$. Then*

$$W_2^2(S_\tau(\mu), \nu) \leq (1 - \tau\lambda)W_2^2(\mu, \nu) - 2\tau(\mathcal{G}(S_\tau(\mu)) - \mathcal{G}(\nu)) - (1 - \tau L)W_2^2(\mu, S_\tau(\mu)), \tag{4.4}$$

for all $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. From the proof of [63, Proposition 8], we can see that there exists a strong Fréchet subgradient of $\mathcal{H}$ at $S_\tau(\mu)$ denoted by $\nabla_W \mathcal{H}(S_\tau(\mu))$ such that

$$\begin{aligned} W_2^2(S_\tau(\mu), \nu) \leq{}& (1 - \tau\lambda)W_2^2(\mu, \nu) - 2\tau(\mathcal{G}(S_\tau(\mu)) - \mathcal{G}(\nu)) \\ &- (1 - \tau L)\|\tau\nabla F + \tau\nabla_W \mathcal{H}(S_\tau(\mu)) \circ X\|_\mu^2, \end{aligned}$$

where $X = T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F)$ and $T_\eta^{\bar{\mu}^+}$ is the optimal transport map between $\eta := (I - \tau\nabla F)_\#(\mu)$ and $\bar{\mu}^+ := J_{\tau,\mathcal{H}}(\eta) = S_\tau(\mu)$. On the other hand, since $(I + \tau\nabla_W \mathcal{H}(S_\tau(\mu)))$ is the optimal transport map between $\bar{\mu}^+$ and η (see for example [63, Lemma 3] or [4]), we have $(I + \tau\nabla_W \mathcal{H}(S_\tau(\mu))) \circ T_\eta^{\bar{\mu}^+} = I$, and

$$\begin{aligned} &\tau\nabla F + \tau\nabla_W \mathcal{H}(S_\tau(\mu)) \circ X \\ &\qquad = \tau\nabla F + (I + \tau\nabla_W \mathcal{H}(S_\tau(\mu))) \circ T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F) - T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F) \qquad (4.5) \\ &\qquad = \tau\nabla F + I - \tau\nabla F - T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F) = I - T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F). \end{aligned}$$

By [63, Lemma 2], $(I - \tau\nabla F)$ is the optimal transport map between μ and η and thus $T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F)$ is a transport map between μ and $S_\tau(\mu)$, which implies $\|I - T_\eta^{\bar{\mu}^+} \circ (I - \tau\nabla F)\|_\mu^2 \geq W_2^2(\mu, S_\tau(\mu))$. Therefore, since $1 - \tau L > 0$, we obtain (4.4). □

In our analysis, the input μ will be only an approximation of the previous iterate, and thus, we would like to remove the condition $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ in the previous lemma. This becomes crucial while considering optimization schemes used to approximate the output of the JKO operator, since the approximated output is usually a discrete measure (if not parametrized otherwise). We thus state and prove now the lemma in its more general form.

Lemma 4.2. *Let $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and F satisfying (A1)-(A3). Let $S_\tau := J_{\tau,\mathcal{H}} \circ (I - \tau\nabla F)_\#$ with $\tau < 1/L$. Then*

$$W_2^2(S_\tau(\mu), \nu) \leq (1 - \tau\lambda)W_2^2(\mu, \nu) - 2\tau(\mathcal{G}(S_\tau(\mu)) - \mathcal{G}(\nu)) - (1 - \tau L)W_2^2(\mu, S_\tau(\mu)), \quad (4.6)$$

for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. We define $\{\mu_n^r\}_n$ as a sequence of regular measures such that $W_2(\mu_n^r, \mu) \to 0$, which always exists (for example smoothing with a convolution). Then, the continuity of $J_{\tau,\mathcal{H}}$ (see Theorem 2.13 (iii)) and the continuity of $(I - \tau\nabla F)_\#$ (which follows by the fact that the operator $I - \tau\nabla F$ is Lipschitz) implies that $W_2(S_\tau(\mu_n^r), S_\tau(\mu)) \to 0$. We have both

$$|W_2(S_\tau(\mu_n^r), \nu) - W_2(S_\tau(\mu), \nu)| \leq W_2(S_\tau(\mu_n^r), S_\tau(\mu)) \to 0,$$

and

$$|W_2(\mu_n^r, S_\tau(\mu_n^r)) - W_2(\mu, S_\tau(\mu))| \leq W_2(\mu_n^r, \mu) + W_2(S_\tau(\mu), S_\tau(\mu_n^r)) \to 0.$$

By Lemma 4.1 we have

$$W_2^2(S_\tau(\mu_n^r), \nu) \leq (1 - \tau\lambda)W_2^2(\mu_n^r, \nu) - 2\tau(\mathcal{G}(S_\tau(\mu_n^r)) - \mathcal{G}(\nu)) - (1 - \tau L)W_2^2(\mu_n^r, S_\tau(\mu_n^r)),$$

and, since $\mathcal{G}$ is lower semicontinuous, we can pass to the limit and obtain (4.6). □

Remark 4.3 (Discrete EVI)**.** *Using the previous lemma, we can prove a new EVI inequality for the sequence $\{\mu_n\}_n$ generated as described in* (4.2)*. In fact, for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ it holds*

$$\begin{aligned} W_2^2(S_{\tau_n}(\mu_n), \nu) \leq (1 - \tau_n\lambda)W_2^2(\mu_n, \nu) &- 2\tau_n(\mathcal{G}(S_{\tau_n}(\mu_n)) - \mathcal{G}(\nu)) \\ &- (1 - \tau_n L)W_2^2(\mu_n, S_{\tau_n}(\mu_n)). \end{aligned} \quad (4.7)$$

This extends the result [63, Proposition 8]. This finer EVI inequality, coming from inequality (4.6)*, is the one we need for our analysis.*

Lemma 4.4. *Suppose that $\{\epsilon_n\}_n$ is a positive summable sequence and the sequence $\{\tau_n\}_n \subset \left(0, \frac{1}{L}\right)$ satisfies $\sup_i \tau_i < \frac{1}{L}$. Then for every $\nu \in \arg\min \mathcal{G}$ the sequences $\{W_2(\mu_n, \nu)\}_n$ and $\{W_2(S_{\tau_n}(\mu_n), \nu)\}_n$ converge and there exists a constant $C > 0$ such that*

$$W_2^2(\mu_{n+1}, \nu) \leq W_2^2(S_{\tau_n}(\mu_n), \nu) + C\epsilon_n. \quad (4.8)$$

It also holds $W_2(S_{\tau_n}(\mu_n), \mu_n) \to 0$, as $n \to +\infty$.

Proof. The first part of the proof is similar to the proof of Lemma 3.1, and therefore omitted. Combining (4.8) and the discrete EVI inequality (4.7), we obtain

$$W_2^2(\mu_{n+1},\nu) - W_2^2(\mu_n,\nu) \leq -(1-\tau_n L)W_2^2(S_{\tau_n}(\mu_n),\mu_n) + \tilde{\epsilon}_n,$$

for all $\nu \in \arg\min \mathcal{G}$, where $\tilde{\epsilon}_n := C\epsilon_n$, for all $n \in \mathbb{N}$, with C the constant in (4.8). Since $\sum_n \tilde{\epsilon}_n = C \cdot \sum_n \epsilon_n < +\infty$ we obtain by summing up

$$\sum_n (1-\tau_n L)W_2^2(S_{\tau_n}(\mu_n),\mu_n) < +\infty. \tag{4.9}$$

Since by hypothesis we have $\sup_i \tau_i < \frac{1}{L}$, it holds in particular $W_2(S_{\tau_n}(\mu_n),\mu_n) \to 0$. □

From now on, we will use the notation $\tilde{\epsilon}_n := C\epsilon_n$, for all $n \in \mathbb{N}$, with C the constant given by the previous lemma.

Lemma 4.5. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\eta := (I - \tau\nabla F)_{\#}\mu$ and $\bar{\mu}^+ := J_{\tau,\mathcal{H}}(\eta)$, with $\tau < \frac{1}{L}$. Then, for every $\bar{\mu} \in \mathcal{P}_2^r(\mathbb{R}^d)$ with $W_2(\bar{\mu},\mu) \leq \epsilon$, we have that*

$$\mathcal{G}(\bar{\mu}^+) - \mathcal{G}(\bar{\mu}) \leq \frac{\epsilon}{\tau}(W_2(\bar{\mu},\eta) + W_2(\bar{\mu}^+,\eta) + \epsilon) \tag{4.10}$$

Proof. Let $\bar{\mu} \in \mathcal{P}_2^r(\mathbb{R}^d)$ such that $W_2(\bar{\mu},\mu) \leq \epsilon$. For every $\bar{\eta} \in \mathcal{P}_2^r(\mathbb{R}^d)$ such that $W_2(\bar{\eta},\eta) \leq \delta$, we have that

$$\begin{aligned}
\mathcal{H}(\bar{\mu}^+) - \mathcal{H}(\bar{\mu}) &\leq \frac{1}{2\tau}W_2^2(\bar{\mu},\eta) - \frac{1}{2\tau}W_2^2(\bar{\mu}^+,\eta) \\
&\leq \frac{1}{2\tau}W_2^2(\bar{\mu},\bar{\eta}) - \frac{1}{2\tau}W_2^2(\bar{\mu}^+,\bar{\eta}) + C_\delta\frac{\delta}{\tau} \\
&= \frac{1}{2\tau}\int \|T_{\bar{\eta}}^{\bar{\mu}}(z) - z\|^2 - \|T_{\bar{\eta}}^{\bar{\mu}^+}(z) - z\|^2 \, d\bar{\eta}(z) + C_\delta\frac{\delta}{\tau} \\
&= \frac{1}{2\tau}\int \|T_{\bar{\eta}}^{\bar{\mu}}(z)\|^2 - \|T_{\bar{\eta}}^{\bar{\mu}^+}(z)\|^2 + 2\langle z, T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\rangle \, d\bar{\eta}(z) + C_\delta\frac{\delta}{\tau},
\end{aligned}$$

where C_δ can be choosen as $C_\delta = W_2(\bar{\mu},\eta) + W_2(\bar{\mu}^+,\eta) + \delta$. On the other hand

$$\begin{aligned}
\mathcal{E}_F(\bar{\mu}^+) - \mathcal{E}_F(\bar{\mu}) &= \int F(T_{\bar{\eta}}^{\bar{\mu}^+}(z)) - F(T_{\bar{\eta}}^{\bar{\mu}}(z)) \, d\bar{\eta}(z) \\
&\leq \int \langle \nabla F(T_{\bar{\eta}}^{\bar{\mu}}(z)), T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\rangle + \frac{L}{2}\|T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\|^2 \, d\bar{\eta}(z) \\
&= \frac{1}{2\tau}\int \Big(-2\langle (I-\tau\nabla F)(T_{\bar{\eta}}^{\bar{\mu}}(z)), T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\rangle \\
&\qquad - 2\langle T_{\bar{\eta}}^{\bar{\mu}}(z), T_{\bar{\eta}}^{\bar{\mu}}(z) - T_{\bar{\eta}}^{\bar{\mu}^+}(z)\rangle + \tau L\|T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\|^2 \Big) \, d\bar{\eta}(z).
\end{aligned}$$

Putting all together leads to

$$\begin{aligned}
\mathcal{G}(\bar{\mu}^+) - \mathcal{G}(\bar{\mu}) \leq \frac{1}{2\tau}\int \Big(& 2\langle z - (I-\tau\nabla F)(T_{\bar{\eta}}^{\bar{\mu}}(z)), T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\rangle \\
& - (1-\tau L)\|T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\|^2 \Big) \, d\bar{\eta}(z) + C_\delta\frac{\delta}{\tau}
\end{aligned} \tag{4.11}$$

Since $\bar{\eta}$ was arbitrary, we can actually choose $\bar{\eta} = (I - \tau\nabla F)_{\#}\bar{\mu}$. With this choice, using the fact that the map $I - \tau\nabla F$ is nonexpansive, we have (see for example [17, Proposition 4.2])

$$W_2(\bar{\eta}, \eta) = W_2((I - \tau\nabla F)_{\#}\bar{\mu}, (I - \tau\nabla F)_{\#}\mu) \leq W_2(\bar{\mu}, \mu) \leq \epsilon,$$

and we can set $\delta = \epsilon$. Defining $\phi = \frac{1}{2}\|\cdot\|^2 - \tau F$, we have $\nabla\phi = I - \tau\nabla F$ and since ∇F is L-Lipschitz and $\tau < \frac{1}{L}$, then $\phi = \frac{1}{2}\|\cdot\|^2 - \tau F$ is strongly convex. This implies that $(I - \tau\nabla F)^{-1}$ is the gradient of a convex function and it is Lipschitz continuous (and thus also in $L^2(\bar{\eta})$). We can therefore apply Theorem 2.1 (ii) and obtain $T_{\bar{\eta}}^{\bar{\mu}} = (I - \tau\nabla F)^{-1}$ and $z - (I - \tau\nabla F)(T_{\bar{\eta}}^{\bar{\mu}}(z)) = 0$ for $\bar{\eta}$-almost every z.

Finally, (4.11) yields

$$\mathcal{G}(\bar{\mu}^+) - \mathcal{G}(\bar{\mu}) \leq 0 - \frac{1 - \tau L}{2\tau} \int \|T_{\bar{\eta}}^{\bar{\mu}^+}(z) - T_{\bar{\eta}}^{\bar{\mu}}(z)\|^2 \, d\bar{\eta}(z) + C_\epsilon \frac{\epsilon}{\tau} \leq C_\epsilon \frac{\epsilon}{\tau}. \tag{4.12}$$

□

Theorem 4.6. *Let $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and F satisfying (A1)-(A3) and suppose* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ with $\sum_{n=0}^{\infty} \epsilon_n < \infty$, $\{\tau_n\}_n \subset \left(0, \frac{1}{L}\right)$ with $\sum_{i=0}^{\infty} \tau_i = \infty$, $\sup_i \tau_i < \frac{1}{L}$ and let $\sigma_n := \sum_{i=0}^{n-1} \tau_i$, for $n \in \mathbb{N}$. Let $\{\mu_n\}_n$ satisfying*

$$W_2(\mu_{n+1}, S_{\tau_n}(\mu_n)) \leq \epsilon_n, \quad \textit{for all } n \in \mathbb{N}.$$

1. *Then, we have*

$$\mathcal{G}(\bar{\beta}_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{as } n \to \infty, \tag{4.13}$$

where $\bar{\beta}_n := S_{\tau_{j_n}}(\mu_{j_n})$ with $j_n = \arg\min_{i=0,\ldots,n-1}\{\mathcal{G}(S_{\tau_i}(\mu_i))\}$, defines the sequence of the best iterates.

2. *If $\sum_{n=0}^{\infty} \frac{\sigma_n}{\tau_n}\epsilon_{n-1} < \infty$, then*

$$\mathcal{G}(S_{\tau_n}(\mu_n)) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{for } n \to \infty,$$

and $\{\mu_n\}_n$ converges with respect to the topology $\tau_{w,2}$ to some $\mu^* \in \arg\min \mathcal{G}$. In particular we have $W_p(\mu_n, \mu^*) \to 0$ for all $p \in [1, 2)$.

Proof. The convergence analysis follows similar steps as the ones depicted in Section 3.1. Summing the EVI (4.7) and using (4.8), we have the first main ingredient, corresponding to (3.7)

$$\begin{aligned}
2\sum_{n=0}^{N-1} \tau_n \mathcal{G}(S_{\tau_n}(\mu_n)) + W_2^2(\mu_N, \nu) &\leq 2\sigma_N \mathcal{G}(\nu) + W_2^2(\mu^0, \nu) \\
&\quad - \sum_{n=0}^{N-1}(1 - \tau_n L) W_2^2(S_{\tau_n}(\mu_n), \mu_n) + \sum_{n=0}^{N-1} \tilde{\epsilon}_n.
\end{aligned} \tag{4.14}$$

From Lemma 4.5 we obtain the second main ingredient, corresponding to (3.10). In fact, for all $n \in \mathbb{N}$, letting $\mu = \mu_n$, $\bar{\mu}^+ = S_{\tau_n}(\mu_n)$ and $\bar{\mu} = S_{\tau_{n-1}}(\mu_{n-1})$ we obtain from (4.10) and the fact that the sequences in play are bounded, that there exists $C > 0$ such that

$$\mathcal{G}(S_{\tau_n}(\mu_n)) \leq \mathcal{G}(S_{\tau_{n-1}}(\mu_{n-1})) + C\frac{\epsilon_{n-1}}{2\tau_n}, \quad \text{for all } n \in \mathbb{N}. \tag{4.15}$$

With these ingredients, it is possible to conclude similarly to Theorem 3.3. □

Remark 4.7. *In the case $\lambda > 0$, using* (4.7) *and* (4.8), *we obtain for all $\nu \in \arg\min \mathcal{G}$ that*

$$W_2^2(\mu_{n+1}, \nu) \leq (1 - \tau_n\lambda)W_2^2(\mu_n, \nu) + \tilde{\epsilon}_n.$$

From this, strong convergence results in W_2 can be achieved. However, since at each iteration the error ϵ_n is committed, we cannot expect linear convergence rates. This motivates us not to treat the case $\lambda > 0$ separately from the case $\lambda = 0$. A similar reasoning applies when assuming strong convexity along generalized geodesics of the functional $\mathcal{H}$, see also Remark 3.4.

Remark 4.8 (Ergodic convergence)**.** *Similar considerations to those in Remark 3.6 apply in this setting as well. In particular, under similar conditions to those in Remark 3.6, the convergence rate in* (4.13) *can also be obtained for the Wasserstein barycenter sequence $\{\bar{\beta}_n\}_n$ formed from the first n elements of the sequence $\{S_{\tau_i}(\mu_i)\}_i$, with weights $\{\frac{\tau_i}{\sigma_n}\}_0^{n-1}$. This result can hold without the additional assumptions required in point 2 of Theorem 4.6.*

With Theorem 4.6 we can generalize the result by Diao, Balasubramanian, Chewi and Salim [34, Theorem 5.3]. In their work they prove weak convergence for the proximal-gradient algorithm but they only work with Gaussians and the so-called Bures-Wasserstein space.

Corollary 4.9. *Let $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and F satisfying (A1)-(A3) and suppose* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\mu_0 \in \mathcal{P}_2(X)$ and $\{\mu_n\}_n$ satisfying*

$$\mu_{n+1} = J_{\tau,\mathcal{H}} \circ (I - \tau\nabla F)_{\#}(\mu_n).$$

Then $\{\mu_n\}_n$ converges with respect to the topology $S_{w,2}$ (and thus also narrowly) to some $\mu^* \in \arg\min \mathcal{G}$.

Proof. We can apply the previous result with $\epsilon_n = 0$ and $\tau_n = \tau > 0$, for all $n \in \mathbb{N}$. □

Theorem 4.10. *Let $\mathcal{H} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and F satisfying (A1)-(A3) and suppose* $\arg\min \mathcal{G} \neq \emptyset$. *Let $\{\epsilon_n\}_n \subset \mathbb{R}_{\geq 0}$ with $\sum_{n=0}^{\infty} \epsilon_n < \infty$, $\{\tau_n\}_n \subset \left(0, \frac{1}{L}\right)$ with $\sum_{i=0}^{\infty} \tau_i = \infty$, $\sup_i \tau_i < \frac{1}{L}$ and let $\sigma_n := \sum_{i=0}^{n-1} \tau_i$, for $n \in \mathbb{N}$. Let $\{\eta_n\}_n$ and $\{\mu_n\}_n$ satisfying $\eta_n = (I - \tau\nabla F)_{\#}(\mu_n)$ and*

$$\mathcal{H}(\mu_{n+1}) + \frac{1}{2\tau}W_2^2(\mu_{n+1}, \eta_n) \leq \mathcal{H}(J_{\tau_n,\mathcal{H}}(\eta_n)) + \frac{1}{2\tau}W_2^2(J_{\tau_n,\mathcal{H}}(\eta_n), \eta_n) + \frac{\epsilon_n^2}{2\tau_n}, \tag{4.16}$$

1. *It holds the rate*
$$\mathcal{G}(\beta_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \text{as } n \to \infty, \tag{4.17}$$
where $\beta_n := \mu_{j_n}$ with $j_n = \arg\min_{i=0,\dots,n-1}\{\mathcal{G}(\mu_i)\}$, defines the sequence of the best iterate.

2. If $\sum_{n=0}^{\infty} \frac{\sigma_n}{\tau_n}\epsilon_n < \infty$, then

$$\mathcal{G}(\mu_n) - \inf \mathcal{G} = O\left(\frac{1}{\sigma_n}\right), \quad \textit{for } n \to \infty,$$

and $\{\mu_n\}_n$ converges with respect to the topology $\tau_{w,2}$ to some $\mu^* \in \arg\min \mathcal{G}$. In particular we have $W_p(\mu_k, \mu^*) \to 0$ for all $p \in [1, 2)$.

Proof. Following the first part of Theorem 3.8, we obtain similarly from the condition (4.16), that

$$W_2(S_{\tau_n}(\mu_n), \mu_{n+1}) = W_2(J_{\tau_n,\mathcal{H}}(\eta_n), \mu_{n+1}) \le \epsilon_n.$$

With this, we can already prove all the results of Theorem 4.6. On the other hand, the inequality (4.16) implies

$$\begin{aligned} 2\tau_n\mathcal{H}(S_{\tau_n}(\mu_n)) &\ge 2\tau_n\mathcal{H}(\mu_{n+1}) + W_2^2(\mu_{n+1}, \eta_n) - W_2^2(S_{\tau_n}(\mu_n), \eta_n) - \epsilon_k^2 \\ &\ge 2\tau_n\mathcal{H}(\mu_{n+1}) - \left(W_2(\mu_{n+1}, \eta_n) + W_2(J_\tau(\eta_n), \eta_n)\right) W_2(\mu_{n+1}, S_{\tau_n}(\mu_n)) - \epsilon_n^2 \\ &\ge 2\tau_n\mathcal{H}(\mu_{n+1}) - c_1\epsilon_n - \epsilon_n^2, \end{aligned}$$

for some $c_1 > 0$. Setting $\bar{\mu}_{n+1} := S_{\tau_n}(\mu_n)$ we also have

$$\begin{aligned} \mathcal{E}_F(\bar{\mu}_{n+1}) &= \int F(x)\, d\bar{\mu}_{n+1}(x) \\ &\ge \int F\left(T_{\bar{\mu}_{n+1}}^{\mu_{n+1}}(x)\right) + \langle \nabla F\left(T_{\bar{\mu}_{n+1}}^{\mu_{n+1}}(x)\right), x - T_{\bar{\mu}_{n+1}}^{\mu_{n+1}}(x)\rangle \, d\bar{\mu}_{n+1}(x) \\ &\ge \mathcal{E}_F(\mu_{n+1}) - \left(\int \left\|\nabla F(x')\right\|^2 \, d\mu_{n+1}(x')\right)^{\frac{1}{2}} W_2(\mu_{n+1}, S_{\tau_n}(\mu_n)). \end{aligned}$$

By noticing that $\|\nabla F(x')\|^2 \le 2\|\nabla F(x') - \nabla F(0)\|^2 + 2\|\nabla F(0)\|^2 \le L^2\|x'\|^2 + 2\|\nabla F(0)\|^2$, and $\{\mu_n\}_n \subset \mathcal{P}_2(\mathbb{R}^d)$ is bounded, and $\tau_n < \frac{1}{L}$, there exists a constant $c_2 > 0$ such that

$$2\tau_n\mathcal{E}_F(\bar{\mu}_{n+1}) \ge 2\tau_n\mathcal{E}_F(\mu_{n+1}) - c_2\epsilon_n$$

and thus

$$2\tau_n\mathcal{G}(S_{\tau_n}(\mu_n)) \ge 2\tau_n\mathcal{G}(\mu_{n+1}) - c_1\epsilon_n - \epsilon_n^2 - c_2\epsilon_n \tag{4.18}$$

Combining this with (4.14), we obtain

$$2\sum_{n=0}^{N-1} \tau_n\mathcal{G}(\mu_{n+1}) \le 2\sigma_N\mathcal{G}(\nu) + W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} \tilde{\epsilon}_n + (c_1 + c_2)\sum_{n=0}^{N-1} \epsilon_n + \sum_{n=0}^{N-1} \epsilon_n^2. \tag{4.19}$$

Since by (4.16) we also have $\mu_n \in \mathcal{P}_2^r(\mathbb{R}^d)$ for all $n \in \mathbb{N}$, we can use Lemma 4.5 with $\bar{\mu}^+ = S_{\tau_n}(\mu_n)$, $\bar{\mu} = \mu_n$ and obtain

$$\mathcal{G}(S_{\tau_n}(\mu_n)) \le \mathcal{G}(\mu_n) + c_3\frac{\epsilon_n}{2\tau_n} \tag{4.20}$$

for some $c_3 > 0$. Combining this with (4.18), we arrive to

$$\mathcal{G}(\mu_{n+1}) \le \mathcal{G}(\mu_n) + (c_1 + c_2 + c_3)\frac{\epsilon_n}{2\tau_n} + \frac{\epsilon_n^2}{2\tau_n}.$$

We denote by $C = c_1 + c_2 + c_3 + \max_i \epsilon_i$, multiply by σ_n, and obtain

$$(\sigma_{n+1} - \tau_n)\mathcal{G}(\mu_{n+1}) - \sigma_n \mathcal{G}(\mu_n) \leq C \frac{\sigma_n}{2\tau_n} \epsilon_n. \tag{4.21}$$

Summing up from 0 to $N-1$ and recalling that $\sigma_0 = 0$, we have

$$\sigma_N \mathcal{G}(\mu_N) - C \sum_{n=0}^{N-1} \frac{\sigma_n}{2\tau_n} \epsilon_n \leq \sum_{n=0}^{N-1} \tau_n \mathcal{G}(\mu_{n+1}). \tag{4.22}$$

which combined with (4.19) leads to

$$\mathcal{G}(\mu_N) - \mathcal{G}(\nu) \leq \frac{1}{2\sigma_N} \left(W_2^2(\mu^0, \nu) + \sum_{n=0}^{N-1} \left(\tilde{\epsilon}_n + (c_1 + c_2)\epsilon_n + \epsilon_n^2 + C \frac{\sigma_n}{\tau_n} \epsilon_n \right) \right),$$

which concludes the proof. □

Conclusions

In this paper, we studied the convergence properties of inexact Jordan-Kinderlehrer-Otto (JKO) schemes and proximal-gradient algorithms in Wasserstein spaces. We focused on settings where obtaining exact solutions to iterative minimization problems is impractical and introduced controlled approximations for both the Wasserstein distance and the energy functionals. We provided rigorous convergence guarantees, demonstrating that weak convergence remains attainable in the presence of inexact computations. Additionally, we extended our analysis to proximal-gradient algorithms. Our findings lay the groundwork for broader applicability of these schemes in practical settings. We also incorporated the flexibility of varying stepsizes, leading to new convergence insights. This study also raises several compelling questions. Notably, it highlights the importance of quantifying the approximation behavior of existing methods, such as ALG2 in [9], when applied to solve (1.1). Future research may pursue this direction by analyzing a range of existing algorithms and proposing new ones for solving (1.1) or the associated saddle point problem derived from the Benamou–Brenier formulation, with quantitative approximation results. This analysis can be complemented by estimates concerning approximations coming from regularized formulations as in [20, 50–52]. Further error estimates may also be developed to account for discretization of measures and, in the Benamou–Brenier case, temporal discretization.

The regularity property (A3) seems essential for the analysis carried out in [63] as well as for our current analysis, but we conjecture that it might be weakened or removed altogether by employing the subdifferential in [4, Theorem 10.3.6]. Investigating this possibility is a promising avenue for future work.

Acknowledgments

S.D.M, E.N. and S.V. are supported by the MUR Excellence Department Project awarded to the department of mathematics at the University of Genoa, CUP D33C23001110001, by the US Air Force Office of Scientific Research (FA8655-22-1-7034). S.V. acknowledges the support of the European Commission (grant TraDE-OPT 861137). S.D.M. and S.V. acknowledge the support of the Ministry of Education, University and Research (grant BAC FAIR PE00000013 funded by the EU - NGEU) and the financial support of the PRIN 202244A7YL. S.D.M., E.N. and S. V. are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains.

References

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

[2] Ya. I. Alber, Regina Burachik, and Alfred Iusem. A proximal point method for nonsmooth convex optimization problems in banach spaces. In *Abstract and Applied Analysis*, volume 2, pages 97–120, 1997.

[3] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*, volume 169 of *UNITEXT*. Springer, 2024.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2nd edition, 2008.

[5] Alfred Auslender. Numerical methods for nondifferentiable convex optimization. In *Nonlinear Anal ysis and Optimization, Mathematical Programming Studies*, pages 102––126. 1987.

[6] Francesca Bartolucci, Marcello Carioni, José A. Iglesias, Yury Korolev, Emanuele Naldi, and Stefano Vigogna. A lipschitz spaces view of infinitely wide shallow neural networks. *arXiv preprint*, 2024. arXiv:2410.14591.

[7] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017.

[8] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[9] Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54:1–17, 2016.

[10] Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Conference on Learning Theory*, pages 1777–1798, 2018.

[11] Malcolm Bowles and Martial Agueh. Weak solutions to a fractional Fokker-–Planck equation via splitting and Wasserstein gradient flow. *Applied Mathematics Letters*, 42:30–35, 2015.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. *Distributed optimization and statistical learning via the alternating direction method of multipliers*, volume 3. Now Publishers Inc., 2011.

[13] Kristian Bredies, Enis Chenchene, and Emanuele Naldi. Graph and distributed extensions of the Douglas–Rachford method. *SIAM Journal on Optimization*, 34(2):1569–1594, 2024.

[14] Kristian Bredies and Hongpeng Sun. Preconditioned Douglas–Rachford splitting methods for convex-concave saddle-point problems. *SIAM Journal on Numerical Analysis*, 53(1):421–444, 2015.

[15] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[16] Regina Burachik and Benar Fux Svaiter. A relative error tolerance for a family of generalized proximal point methods. *Mathematics and Operations Research*, 26:816––831, 2001.

[17] Arian Bërdëllima and Gabriele Steidl. Quasi α-firmly nonexpansive mappings in Wasserstein spaces. *Fixed Point Theory*, 26(1):37–56, 2025.

[18] Eric A. Carlen and Katy Craig. Contraction of the proximal map and generalized convexity of the Moreau-Yosida regularization in the 2-Wasserstein metric. *Mathematics and Mechanics of Complex Systems*, 1(1):33–65, 2013.

[19] Guillaume Carlier, Alex Delalande, and Quentin Mérigot. Quantitative stability of barycenters in the Wasserstein space. *Probability Theory and Related Fields*, 188(3):1257–1286, Apr 2024.

[20] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[21] Guillaume Carlier and Maxime Laborde. A splitting method for nonlinear diffusions with nonlocal, nonpotential drifts. *Nonlinear Analysis: Theory, Methods & Applications*, 150:1–18, 2017.

[22] Giulia Cavagnari, Giuseppe Savaré, and Giacomo Enrico Sodini. A Lagrangian approach to totally dissipative evolutions in Wasserstein spaces, 2023.

[23] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[24] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211, 2018.

[25] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

[26] Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004.

[27] Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[28] Roberto Cominetti. Coupling the proximal point algorithm with approximation methods. *Journal of Optimization Theory and Applications*, 95:581–600, 1997.

[29] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. Curran Associates, Inc., 2013.

[30] Damek Davis. Convergence rate analysis of the forward-Douglas–Rachford splitting scheme. *SIAM Journal on Optimization*, 25(3):1760–1786, 2015.

[31] Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 115–163. Springer International Publishing, 2016.

[32] Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. BV estimates in optimal transportation and applications. *Archive for Rational Mechanics and Analysis*, 219(2):829–860, 2016.

[33] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, Aug 2014.

[34] Michael Diao, Krishnakumar Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space, 2023.

[35] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2018.

[36] Jonathan Eckstein. Approximate iterations in bregman-function-based proximal algorithms. *Mathematical Programming*, 83:113–123, 1998.

[37] Jonathan Eckstein and Dimitri P. Bertsekas. Douglas-Rachford splitting methods in convex programming. *Mathematical Programming*, 55(1-3):293–318, 1992.

[38] Michel Fortin and Roland Glowinski. On decomposition-coordination methods using an augmented lagrangian. In *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, pages 97–146. North-Holland, Amsterdam, 1983.

[39] Daniel Gabay. Applications of the method of multipliers to variational inequalities. In *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, pages 299–331. North-Holland, Amsterdam, 1983.

[40] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[41] E. De Giorgi. New problems on minimizing movements. In *Boundary Value Problems for PDE and Applications*, pages 81–98. Masson, 1993.

[42] Roland Glowinski and Patrick Le Tallec. Augmented lagrangian methods for the solution of variational problems. Mrc technical summary report #2965, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1987.

[43] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[44] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM journal on control and optimization*, 29(2):403–419, 1991.

[45] Howard Heaton, Samy Wu Fung, and Stanley Osher. Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs. *Communications on Applied Mathematics and Computation*, 6(2):790–810, 2024.

[46] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

[47] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[48] Paul Knopp and Richard Sinkhorn. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343 – 348, 1967.

[49] M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, May 1984.

[50] Wonjun Lee, Li Wang, and Wuchen Li. Deep JKO: Time-implicit particle methods for general nonlinear gradient flows. *Journal of Computational Physics*, 514:113187, 2024.

[51] Wuchen Li, Siting Liu, and Stanley Osher. A kernel formula for regularized Wasserstein proximal operators. *Research in the Mathematical Sciences*, 10(4):43, 2023.

[52] Wuchen Li, Jianfeng Lu, and Li Wang. Fisher information regularization schemes for Wasserstein gradient flows. *Journal of Computational Physics*, 416:109449, 2020.

[53] Alex Tong Lin, Wuchen Li, Stanley Osher, and Guido Montúfar. Wasserstein proximal of GANs. In *Geometric Science of Information*, pages 524–533, Cham, 2021. Springer International Publishing.

[54] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

[55] Bernard Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B*, 274:163–165, 1972.

[56] Bertrand Maury, Aude Roudneff-Chupin, Filippo Santambrogio, and Juliette Venel. Handling congestion in crowd motion modeling. *Networks and Heterogeneous Media*, pages 485–519, 2011.

[57] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 1–77, 2019.

[58] Emanuele Naldi and Giuseppe Savaré. Weak topology and opial property in Wasserstein spaces, with applications to gradient flows and proximal point algorithms of geodesically convex functionals. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur.*, 32(4):725–750, 2021.

[59] Stanley Osher, Heather Heaton, and Shingyu Wu Fung. A Hamilton-Jacobi-based proximal operator. *Proceedings of the National Academy of Sciences of the United States of America*, 120(14):e2220469120, 2023.

[60] Michael J.D. Powell. A method for nonlinear constraints in minimization problems. pages 283–298, 1969.

[61] Pierre H. Richemond and Brendan Maginnis. On Wasserstein reinforcement learning and the Fokker–Planck equation. *arXiv preprint*, 2017. arXiv:1712.07185.

[62] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[63] Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pages 12356–12366. Curran Associates, Inc., 2020.

[64] Saverio Salzo, Silvia Villa, et al. Inexact and accelerated proximal point algorithms. *Journal of Convex analysis*, 19(4):1167–1192, 2012.

[65] Mikhail Solodov and Svaiter Benar Fux. Error bounds for proximal point subproblems and associated inexact proximal point algorithms. *Mathematical Programming*, 88:371–389, 2000.

[66] Mikhail Solodov and Benar Fux Svaiter. A unified framework for some inexact proximal point algorithms. *Numerical functional analysis and optimization*, 22(7-8):1013–1035, 2001.

[67] Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, 2013.

[68] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[69] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 2018.

[70] Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as Wasserstein gradient flows. In *International Conference on Machine Learning*, pages 5737–5746, 2018.