Wasserstein Proximal Gradient

Adil Salim[*]   Anna Korba[†]   Giulia Luise[‡]

February 11, 2020

Abstract

We consider the task of sampling from a log-concave probability distribution. This target distribution can be seen as a minimizer of the relative entropy functional defined on the space of probability distributions. The relative entropy can be decomposed as the sum of a functional called the potential energy, assumed to be smooth, and a nonsmooth functional called the entropy. We adopt a Forward Backward (FB) Euler scheme for the discretization of the gradient flow of the relative entropy. This FB algorithm can be seen as a proximal gradient algorithm to minimize the relative entropy over the space of probability measures. Using techniques from convex optimization and optimal transport, we provide a non-asymptotic analysis of the FB algorithm. The convergence rate of the FB algorithm matches the convergence rate of the classical proximal gradient algorithm in Euclidean spaces. The practical implementation of the FB algorithm can be challenging. In practice, the user may choose to discretize the space and work with empirical measures. In this case, we provide a closed form formula for the proximity operator of the entropy.

1   Introduction

The task of sampling with respect to some target distribution $\mu_\star$ has become a pre-requesite for many inference procedures. We consider the case where $\mu_\star$ admits a density proportional to $\exp(-F)$ with respect to Lebesgue measure over $\mathcal{X} = \mathbb{R}^d$ (we shall write $\mu_\star \propto \exp(-F)$). The function F, referred to as the *potential function*, is assumed convex and smooth. Our sampling problem can be approached from an optimization point of view. Indeed, the target distribution $\mu_\star$ is the solution to the optimization problem defined on the set $\mathcal{P}_2(\mathcal{X})$ of probability measures μ such that $\int \|x\|^2 d\mu(x) < \infty$ by

$$\min_{\mu \in \mathcal{P}_2(\mathcal{X})} \mathrm{KL}(\mu|\mu_\star), \tag{1}$$

where KL denotes the Kullback-Leibler divergence, or relative entropy.[1]

---

[*]`adil.salim@kaust.edu.sa`, Visual Computing Center, KAUST

[†]`a.korba@ucl.ac.uk`, Gatsby Computational Neuroscience Unit, University College London

[‡]`g.luise16@ucl.ac.uk`, Computer Science Department, University College London

[1]$\mu_\star \in \mathcal{P}_2(\mathcal{X})$, see [8, Lemma 2.2.1].

Problem (1) can be solved in continuous time by the gradient flow of the relative entropy [1, 19], a family of functions $t \mapsto \mu(t)$ defined over $(0, +\infty)$ and converging to $\mu_\star$. In discrete time, some sampling algorithms can be seen as discretization schemes of the gradient flow of the relative entropy [6, 15, 31]. For example, Langevin algorithm [14, 16] can be seen as an *inexact* gradient descent algorithm to solve Problem (1) over $\mathcal{P}_2(\mathcal{X})$. Because of this inexactness, Langevin algorithm converges at the rate $\mathcal{O}(1/\sqrt{n})$ (where n is the number of iterations), whereas gradient descent in Euclidean spaces converges at the rate $\mathcal{O}(1/n)$. However, exact discretization schemes of the gradient flow of the entropy exist and are reviewed in [31]. These schemes typically rely on iterative applications of the JKO operator, a map from $\mathcal{P}_2(\mathcal{X})$ to $\mathcal{P}_2(\mathcal{X})$. The JKO operator was introduced [19] to construct the gradient flow as the limit of piecewise constant functions with values in $\mathcal{P}_2(\mathcal{X})$. Indeed, the JKO operator is similar to the proximity operator in optimization, a map from $\mathcal{X}$ to $\mathcal{X}$ [4]. Many nonsmooth optimization algorithms rely on iterative applications of the proximity operator. Moreover, the proximity operator can also be used to construct continuous time flows as a limit of piecewise constant functions, see [7, 26].

Our first contribution is to study the Forward Backward (FB) discretization scheme of the gradient flow of the relative entropy [31, Section 4.1] as an optimization algorithm to solve problem (1). We show convergence rates for the FB scheme similar to the proximal gradient algorithm in Euclidean spaces. As [31], we observe that the relative entropy is the sum of a smooth and a nonsmooth functionals, respectively the potential energy and the entropy. The FB scheme runs alternatively a forward method for the potential energy and a backward method for the entropy. The FB scheme can be seen as an *exact* proximal gradient algorithm to solve Problem (1) over $\mathcal{P}_2(\mathcal{X})$ [31, Section 4.1]. Our first contribution is to provide a non-asymptotic analysis of the FB scheme. We show that the FB scheme converges at the rate $\mathcal{O}(1/n)$, and linearly if F is strongly convex. These rates match the convergence rate of the proximal gradient algorithm in Euclidean spaces. Our analysis relies on the theory of gradient flows in the space of probability measures [1]. The main challenge in the convergence proof is to address the non convexity of the entropy functional. Indeed, the entropy functional satisfies a property weaker than convexity, called generalized geodesic convexity. Identifying the optimal transport maps between the iterates of the FB scheme, we profit from this geodesic convexity of the entropy to establish a discrete Evolution Variational Inequality (EVI) for the FB scheme, similarly to the analysis of Langevin algorithm in [15]. The convergence rates are corollaries of the EVI.

The FB scheme is more challenging to implement than Langevin algorithm because of the backward (or proximal, or JKO) step. The practical implementation of the JKO operator has been considered in several papers, see [28, Section 4.8]. In the specific case of the JKO of the entropy functional, [31, Section G.1] provides a closed form formula for the gaussian case.

Our second contribution is to provide a closed form formula for the JKO of the entropy between discrete measures. When the space $\mathcal{X}$ is discretized, the JKO of the entropy has to be redefined to fit the situation. Using a duality approach similar to the one used for regularized optimal transport [13], we are able to give a closed form formula for the JKO of the entropy.

We finally mention two lines of works using proximal techniques in settings close

(but different) from ours. Proximity operators have been used in Langevin algorithm to sample from target distributions with composite potentials [6, 15, 27, 31]. Moreover, the space $\mathcal{P}_2(\mathcal{X})$ can be interpreted as a Riemannian space [23, 25, 30] and the proximal gradient algorithm over finite dimensional Riemannian space have recently been investigated in [11, 18].

The remainder is organized as follows. In Section 2 we provide some background knowledge in optimal transport and gradient flows. In Section 3, we introduce the Forward Backward Euler discretization scheme. This FB scheme is studied as an optimization algorithm in Section 4. We provide non-asymptotic rates for the resolution of (1). Then, a closed form formula for the JKO of the entropy between discrete measures is given in Section 5. Finally, we conclude in Section 6. The convergence proofs are postponed to the appendix.

2 Preliminaries

In this section, we introduce the notations and recall fundamental definitions and properties on optimal transport and gradient flows that will be used throughout the paper.

2.1 Notations

In the sequel, $\mathcal{P}_2(\mathcal{X})$ is the space of probability measures μ on $\mathcal{X}$ with finite second order moment. Denote $\mathcal{B}(\mathcal{X})$ the Borelian σ-field over $\mathcal{X}$. For any $\mu \in \mathcal{P}_2(\mathcal{X})$, $L^2(\mu)$ is the space of functions $f : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that $\int \|f\|^2 d\mu < \infty$. Note that the identity map I is an element of $L^2(\mu)$. For any $\mu \in \mathcal{P}_2(\mathcal{X})$, 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 Leb the Lebesgue measure over $\mathcal{X}$. The set of regular distributions of the Wasserstein space is denoted $\mathcal{P}_2^r(\mathcal{X}) := \{\mu \in \mathcal{P}_2(\mathcal{X}), \ \mu \ll Leb\}$. If $f, g : \mathcal{X} \to \mathcal{X}$, the composition $f \circ g$ of g by f is sometimes denoted $f(g)$.

2.2 Optimal transport

For every measurable map T defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and for every $\mu \in \mathcal{P}_2(\mathcal{X})$, we denote $T_\#\mu$ the pushforward measure of μ by T characterized by the transfer lemma:

$$\int \phi(y)dT_\#\mu(y) = \int \phi(T(x))d\mu(x) \quad \text{for any measurable and bounded function } \phi. \tag{2}$$

Consider the 2-Wasserstein distance defined for every $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$ by

$$W^2(\mu, \nu) := \inf_{\upsilon \in \Gamma(\mu,\nu)} \int \|x - y\|^2 d\upsilon(x, y), \tag{3}$$

where $\Gamma(\mu, \nu)$ is the set of couplings between μ and ν [30], *i.e.* the set of nonnegative measures υ over $\mathcal{X} \times \mathcal{X}$ such that $P_\#\upsilon = \mu$ (resp. $Q_\#\upsilon = \nu$) where $P : (x, y) \mapsto x$ (resp. $Q : (x, y) \mapsto x$).

We now recall the well-known Brenier theorem [9], [1, Section 6.2.3].

Theorem 2.1 *Let $\mu \in \mathcal{P}_2^r(\mathcal{X})$ and $\nu \in \mathcal{P}_2(\mathcal{X})$. Then,*

1. *There exists an unique minimizer υ of* (3). *Besides, there exists an uniquely determined μ-almost everywhere (a.e.) map $T_\mu^\nu : \mathcal{X} \to \mathcal{X}$ such that $\upsilon = (I, T_\mu^\nu)_{\#}\mu$ where $(I, T_\mu^\nu) : (x, y) \mapsto (x, T_\mu^\nu(x))$. Finally, there exists a convex function $f : \mathcal{X} \to \mathbb{R}$ such that $T_\mu^\nu = \nabla f$ μ-a.e.*

2. *As a corollary,*

$$W^2(\mu, \nu) = \int \|x - \nabla f(x)\|^2 d\mu(x) = \inf_{T_{\#}\mu = \nu} \int \|x - T(x)\|^2 d\mu(x). \quad (4)$$

3. *If $g : \mathcal{X} \to \mathbb{R}$ is convex, then ∇g is well defined μ-a.e. and if $\nu = \nabla g_{\#}\mu$, then $T_\mu^\nu = \nabla g$ μ-a.e.*

4. *If $\nu \in \mathcal{P}_2^r(\mathcal{X})$, then $T_\nu^\mu \circ T_\mu^\nu = I$ μ-a.e. and $T_\mu^\nu \circ T_\nu^\mu = I$ ν-a.e.*

Under the assumptions of Theorem 2.1, the map T_μ^ν is called the optimal transport (OT) map from μ to ν. In this paper, as it is commonly the case in the literature, we may refer to the space of probability distributions $\mathcal{P}_2(\mathcal{X})$ equipped with the 2-Wasserstein distance as the Wasserstein space.

2.3 Review of Gradient Flows and their discretizations

2.3.1 In an Euclidean space

Assume that $\mathcal{X}$ is an Euclidean space, consider a proper lower semi-continuous function $G : \mathcal{X} \to (-\infty, +\infty]$ and denote $D(G) = \{x \in \mathcal{X}, \ G(x) < \infty\}$ its domain. We assume that G is convex, *i.e.*, for every $x, z \in \mathcal{X}$ and for every $\varepsilon \in [0, 1]$, we have:

$$G(\varepsilon z + (1 - \varepsilon)x) \le \varepsilon G(z) + (1 - \varepsilon)G(x). \quad (5)$$

Given $x \in \mathcal{X}$, $y \in \mathcal{X}$ is a subgradient of G at x if for every $z \in \mathcal{X}$,

$$G(x) + \langle y, z - x \rangle \le G(z).$$

The (possibly empty) set of subgradients of G at x is denoted $\partial G(x)$, and the map $x \mapsto \partial G(x)$ is called the subdifferential. If G is differentiable at x, then $\partial G(x) = \{\nabla G(x)\}$ where $\nabla G(x)$ is the gradient of G at x. The subdifferential of the convex function G allows to define the gradient flow of G: for every initial condition $\mathsf{x}(0) = a$, there exists an unique absolutely continuous function $\mathsf{x} : [0, +\infty) \to \mathcal{X}$ solution to the differential inclusion [10, Th. 3.1], [26, Th. 2.7]:

$$\mathsf{x}'(t) \in \partial G(\mathsf{x}(t)). \quad (6)$$

One can check that the gradient flow of G is also characterized by the following system of Evolution Variational Inequalities (EVI) :

$$\forall z \in D(G), \quad \frac{d}{dt}\|\mathsf{x}(t) - z\|^2 \le -2\left(G(\mathsf{x}(t)) - G(z)\right).$$

In contrast to (6), the former characterization allows to define the gradient flow without using the notion of subdifferential, a property that can be practical in nonsmooth settings. Moreover, the non-asymptotic analysis of discretized gradient flows in the optimization literature often relies on discrete versions of the EVI.

The existence of Gradient Flows can be established as the limit of a proximal scheme [26, Th. 2.14], [7, Th. 5.1] when the step-size $\gamma \to 0$. Defining the proximity operator of G as:

$$\text{prox}_{\gamma G}(x) := \underset{y \in \mathcal{X}}{\arg\min} \, G(y) + \frac{1}{2\gamma}\|y - x\|^2, \tag{7}$$

the proximal scheme is written

$$x_{n+1} = \text{prox}_{\gamma G}(x_n), \tag{8}$$

which corresponds to the proximal point algorithm to minimize the function G, see [22]. The proximal scheme can be seen as a *Backward Euler discretization* of the gradient flow. Indeed, writing the first order conditions of (8), we have

$$x_{n+1} \in x_n - \gamma\partial G(x_{n+1}), \quad \text{or equivalently} \quad \frac{x_{n+1} - x_n}{\gamma} \in -\partial G(x_{n+1}).$$

Hence, each iteration of the proximal scheme requires solving an equation which can be intractable in many cases. The *Forward Euler scheme* is a more tractable integrator of the gradient flow of G, but is less stable and requires the differentiability of G. Under this assumption, this scheme is written

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla G(x_n) \quad \text{or equivalently} \quad x_{n+1} = x_n - \gamma\nabla G(x_n), \tag{9}$$

which corresponds to the well-known gradient descent algorithm to minimize the function G. Consider now the case where the function G can be decomposed as $G = F + H$, where F is convex and smooth and H is convex and nonsmooth. To integrate the gradient flow of $G = F + H$, another approach is to use the Forward and the Backward Euler schemes for the smooth term and nonsmooth term respectively [26]. This approach is also motivated by the fact that in many situations, the function H is simple enough to implement its proximity operator $\text{prox}_{\gamma H}$. If $G = F + H$, the Forward Backward Euler scheme is written

$$\frac{x_{n+1} - x_n}{\gamma} \in -\nabla F(x_n) - \partial H(x_{n+1}). \tag{10}$$

Recalling the definition of the proximity operator, this scheme can be rewritten as

$$x_{n+1} = \text{prox}_{\gamma H}(x_n - \gamma\nabla F(x_n)), \tag{11}$$

which corresponds to the proximal gradient algorithm to minimize the composite function G.

2.3.2 In the Wasserstein space

Consider a proper lower semi continuous functional $\mathcal{G} : \mathcal{P}_2(\mathcal{X}) \rightarrow (-\infty, +\infty]$ and denote $D(\mathcal{G}) = \{\mu \in \mathcal{P}_2(\mathcal{X}), \mathcal{G}(\mu) < \infty\}$ its domain. We assume that $\mathcal{G}$ is convex along generalized geodesics defined by the 2-Wasserstein distance [1, Chap. 9], *i.e.* for every $\mu, \pi \in \mathcal{P}_2(\mathcal{X}), \nu \in \mathcal{P}_2^r(\mathcal{X})$ and for every $\varepsilon \in [0, 1]$,

$$\mathcal{G}((\varepsilon T_\nu^\pi + (1 - \varepsilon)T_\nu^\mu)_\# \nu) \leq \varepsilon \mathcal{G}(\pi) + (1 - \varepsilon)\mathcal{G}(\mu). \tag{12}$$

where T_ν^π and T_ν^μ are the optimal transport maps from ν to π and from ν to μ respectively. Given $\mu \in \mathcal{P}_2(\mathcal{X}), \xi \in L^2(\mu)$ is a strong Fréchet subgradient of $\mathcal{G}$ at μ [1, Chap. 10] if for every $\phi \in L^2(\mu)$,

$$\mathcal{G}(\mu) + \varepsilon\langle \xi, \phi \rangle_\mu + o(\varepsilon) \leq \mathcal{G}((I + \varepsilon\phi)_\# \mu).$$

The (possibly empty) set of strong Fréchet subgradients of $\mathcal{G}$ at μ is denoted $\partial \mathcal{G}(\mu)$. The map $\mu \mapsto \partial \mathcal{G}(\mu)$ is called the strong Fréchet subdifferential. Conveniently, a subdifferential notion similar to the strong Fréchet subdifferential enables to define the gradient flow of the functional $\mathcal{G}$ [1, Chap. 11]. However in the nonsmooth setting that will be considered in this paper, the characterization of gradient flows through EVI will be more practical. The gradient flow of $\mathcal{G}$ is the solution of the following system of Evolution Variational Inequalities (EVI) [1, Th. 11.1.4]:

$$\forall \pi \in D(\mathcal{G}), \quad \frac{d}{dt}W^2(\mu(t), \pi) \leq -2\left(\mathcal{G}(\mu(t)) - \mathcal{G}(\pi)\right).$$

We shall perform a non-asymptotic analysis of a discretized gradient flow scheme to minimize the functional $\mathcal{G}$. Our approach is to prove a discrete EVI for this scheme.

The existence of gradient flows can be established as the limit of a minimizing movement scheme [1, Th. 11.2.1], [19]. Defining the JKO operator of $\mathcal{G}$ as:

$$\text{JKO}_{\gamma\mathcal{G}}(\mu) := \underset{\nu \in \mathcal{P}_2(\mathcal{X})}{\arg\min}\, \mathcal{G}(\nu) + \frac{1}{2\gamma}W^2(\nu, \mu), \tag{13}$$

the JKO scheme is written

$$\mu_{n+1} = \text{JKO}_{\gamma\mathcal{G}}(\mu_n).$$

The JKO operator can be seen as a proximity operator by replacing the Wasserstein distance by the Euclidean distance. Moreover, the JKO scheme can be seen as a Backward Euler discretization of the gradient flow. More precisely, under some assumptions on the functional $\mathcal{G}$, using [1, Lemma 10.1.2]

$$\frac{T_{\mu_{n+1}}^{\mu_n} - I}{\gamma} \in \partial \mathcal{G}(\mu_{n+1}).$$

Since, using Brenier's theorem, $T_{\mu_{n+1}}^{\mu_n} \circ T_{\mu_n}^{\mu_{n+1}} = I$, there exists a strong Fréchet subgradient of $\mathcal{G}$ at μ_{n+1} denoted $\nabla_W \mathcal{G}(\mu_{n+1})$ such that

$$\mu_{n+1} = \left(I - \gamma\nabla_W \mathcal{G}(\mu_{n+1}) \circ T_{\mu_n}^{\mu_{n+1}}\right)_\# \mu_n.$$

Each iteration of the JKO scheme thus requires the minimization of a function which can be intractable in many cases. As previously, the Forward Euler scheme is more tractable and enjoys additionally a simpler geometrical interpretation, see *e.g.* [31, Section 3.1.3]. Assume $\partial \mathcal{G}(\mu) = \{\nabla \mathcal{G}(\mu)\}$ is a singleton for any $\mu \in D(\mathcal{G})$ (some examples are given [1][Sec. 10.4]). The Forward Euler scheme for the gradient flow of $\mathcal{G}$ is written:

$$\mu_{n+1} = (I - \gamma \nabla \mathcal{G}(\mu_n))_{\#} \mu_n, \tag{14}$$

and corresponds to the iterations of the gradient descent algorithm over the Wasserstein space to minimize $\mathcal{G}$. Although the Wasserstein space is not a Riemannian manifold, it can still be equipped with a Riemannian structure and a Riemannian interpretation [23, 25]. In particular, the Forward Euler scheme can be seen as a Riemannian gradient descent where the exponential map at μ is the map $\phi \mapsto (I + \phi)_{\#}\mu$ defined on $L^2(\mu)$.

3 A Forward Backward Euler scheme for the relative entropy

Given two nonnegative measures μ, π over $\mathcal{P}_2(\mathcal{X})$, the Kullback-Leibler divergence (or relative entropy) of μ from π is defined as:

$$\mathrm{KL}(\mu|\pi) := \int \log\left(\frac{d\mu}{d\pi}(x)\right) d\mu(x)$$

if $\mu \ll \pi$ with density $\frac{d\mu}{d\pi}$, and $\mathrm{KL}(\mu|\pi) := +\infty$ else. Then, for $\mu \in \mathcal{P}_2(X)$, we define the potential energy as:

$$\mathcal{E}_F(\mu) := \int F(x) d\mu(x), \tag{15}$$

and

$$\mathcal{H}(\mu) := \int \log\left(\frac{d\mu}{dLeb}(x)\right) d\mu(x),$$

if $\mu \ll Leb$ with density $\frac{d\mu}{dLeb}$, and $\mathcal{H}(\mu) := +\infty$ else. Throughout this paper, we assume the following on the potential function F: there exists $L, \lambda \geq 0$ such that

- **A1.** F is L-smooth *i.e.* F is differentiable and ∇F is L-Lipschitz continuous; for all $(x, y) \in \mathcal{X}^2$:

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \tag{16}$$

- **A2.** F is λ-strongly convex (we allow $\lambda = 0$); for all $(x, y) \in \mathcal{X}^2$:

$$F(x) \leq F(y) - \langle \nabla F(x), y - x \rangle - \frac{\lambda}{2}\|x - y\|^2. \tag{17}$$

We first recall that the KL divergence can be expressed (up to an additive constant) as a sum of the potential and the entropy functionals.

Lemma 3.1 *[15, Lemma 1] Let $\mu_\star(x) \propto \exp(-F(x))$ be the target distribution. Then, $\mu_\star \in \mathcal{P}_2(\mathcal{X})$, $\mathcal{E}_F(\mu_\star) < \infty$ and $\mathcal{H}(\mu_\star) < \infty$. Moreover, for every $\mu \in \mathcal{P}_2(\mathcal{X})$ satisfying $\mathcal{E}_F(\mu) < \infty$,*

$$\mathrm{KL}(\mu|\mu_\star) = \mathcal{E}_F(\mu) + \mathcal{H}(\mu) - (\mathcal{E}_F(\mu_\star) + \mathcal{H}(\mu_\star)). \tag{18}$$

Based on Lemma (3.1), Problem (1) can be rewritten as the minimization of the potential energy regularized by the entropy:

$$\mu_\star = \arg\min_{\mu \in \mathcal{P}_2(\mathcal{X})} \mathcal{E}_F(\mu) + \mathcal{H}(\mu). \tag{19}$$

Since the target distribution $\mu_\star$ is a minimizer of the composite functional $\mathcal{G} = \mathcal{E}_F + \mathcal{H}$, we use a Forward Backward Euler scheme to integrate the gradient flow of $\mathcal{G}$. As $\mathcal{H}$ is nonsmooth, see [31, Section 3.1.1], this scheme relies on the implementation of the JKO operator of the entropy $\mathrm{JKO}_{\gamma\mathcal{H}}$, treated separately in Section 5. The proposed Forward Backward Euler scheme is written, for $n \geq 0$: [31, Section 4.1]

$$\nu_{n+1} = (I - \gamma\nabla F)_{\#}\mu_n \tag{20}$$
$$\mu_{n+1} = \mathrm{JKO}_{\gamma\mathcal{H}}(\nu_{n+1}). \tag{21}$$

This scheme can be seen as a proximal gradient algorithm over the Wasserstein space to minimize the composite function $\mathcal{G} = \mathcal{E}_F + \mathcal{H}$.

Remark 1 (Comparison with Forward Euler schemes) *The Forward Euler scheme (14) for $\mathcal{G} = \mathcal{E}_F + \mathcal{H}$ was considered in [31, Section 3.1.3]. This scheme is notably difficult to analyze because of the nonsmoothness of $\mathcal{H}$ [31, Section 3.1.1]. Moreover, Stein Variational Gradient Descent (SVGD) algorithm introduced recently [20, 21] can be seen as a Forward Euler scheme under the inner product $\langle \cdot, \cdot \rangle_H$ of a Reproducing Kernel Hilbert Space (RKHS) H dense in $L^2(\mu)$ for every μ. Indeed, SVGD is written*

$$\mu_{n+1} = (I - \gamma P_\mu \nabla\mathcal{G}(\mu_n))_{\#}\mu_n,$$

where $P_\mu : L^2(\mu) \to L^2(\mu)$ is the kernel integral operator defined by $P_\mu f = \int f(x)k(\cdot, x)d\mu(x) \in H$ where k is the kernel of H. The inner product $\langle \cdot, \cdot \rangle_H$ can be seen as the inner product induced by P_μ^{-1} on H. Indeed, for every $f, g \in H$, $\langle f, g \rangle_H = \langle P_\mu^{-1}f, g \rangle_\mu$, see e.g. [3, Section 2.1]. Moreover, $P_\mu\nabla\mathcal{G}(\mu)$ can be seen as the gradient of $\mathcal{G}$ at μ under the inner product $\langle \cdot, \cdot \rangle_H$ since $\langle P_\mu\nabla\mathcal{G}(\mu), g \rangle_H = \langle \nabla\mathcal{G}(\mu), g \rangle_\mu$ for every $g \in H$. Finally, to the best of our knowledge, the non asymptotic performances of SVGD algorithm remain unclear.

In the next section, we study the non asymptotic properties of the FB scheme.

4   Non asymptotic analysis

We consider a fixed step size $\gamma < 1/L$ and a probability distribution $\pi \in \mathcal{P}_2(\mathcal{X})$. Our main result (Proposition 4.7) combines two ingredients: the identification of the optimal transport maps between μ_n, ν_{n+1} and μ_{n+1} (see Equations (20) and (21)), and a proof of a discrete EVI for the proximal gradient algorithm.

4.1 Identification of optimal transport maps

Lemmas 4.1,4.2 identify the optimal transport maps from μ_n to ν_{n+1} and from ν_{n+1} to μ_{n+1} in the Forward Backward Euler scheme, as soon as the step size is sufficiently small. In particular, Lemma 4.2 is a consequence of [1, Lemma 10.1.2].

Lemma 4.1 *Let $\mu \in \mathcal{P}_2^r(\mathcal{X})$ and $\nu = (I - \gamma \nabla F)_{\#}\mu$. Then if $\gamma < 1/L$, the optimal transport map from μ to ν corresponds to*

$$T_\mu^\nu = I - \gamma \nabla F.$$

Moreover, $\nu \in \mathcal{P}_2^r(\mathcal{X})$.

Lemma 4.2 *Let $\nu \in \mathcal{P}_2(\mathcal{X})$. If $\mu = \mathrm{JKO}_{\gamma\mathcal{H}}(\nu)$, then $\mu \in D(\mathcal{H}) \subset \mathcal{P}_2^r(\mathcal{X})$ and the optimal transport map from μ to ν satisfies $T_\mu^\nu \in I + \gamma\partial\mathcal{H}(\mu)$. In other words, there exists a strong Fréchet subgradient at μ denoted $\nabla_W \mathcal{H}(\mu)$ such that*

$$T_\mu^\nu = I + \gamma \nabla_W \mathcal{H}(\mu). \tag{22}$$

Using Lemmas 4.1,4.2, if $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$, then $\mu_n, \nu_n \in \mathcal{P}_2^r(\mathcal{X})$ for every n by induction. This remark allows to consider optimal transport maps from μ_n and ν_n to any $\pi \in \mathcal{P}_2(\mathcal{X})$. The next lemma extends [1, 10.1.1.B] using the generalized geodesic convexity of $\mathcal{H}$ and the strong Fréchet subdifferential of $\mathcal{H}$.

Lemma 4.3 *Let $\nu \in \mathcal{P}_2^r(\mathcal{X})$, $\mu, \pi \in \mathcal{P}_2(\mathcal{X})$ and T_ν^μ, T_ν^π the optimal transport maps from ν to μ and from ν to π respectively. If $\xi \in \partial\mathcal{H}(\mu)$, then*

$$\langle \xi \circ T_\nu^\mu, T_\nu^\pi - T_\nu^\mu \rangle_\nu \leq \mathcal{H}(\pi) - \mathcal{H}(\mu). \tag{23}$$

Equipped with these preliminary results, we can establish a discrete EVI for the FB scheme (20)- (21) in the next subsection.

4.2 A descent lemma

Without using any convexity assumption on F, we first obtain a descent lemma. We denote $Y_{n+1} := T_{\nu_{n+1}}^{\mu_{n+1}}$ the optimal transport map between ν_{n+1} and μ_{n+1} in the Forward Backward Euler scheme (20), (21), and $X_{n+1} := Y_{n+1} \circ (I - \gamma\nabla F)$.

Theorem 4.4 (Descent) *Assume $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$, $\gamma < 1/L$ and **A1**. Then for $n \geq 0$, there exists a strong Fréchet subgradient at μ_{n+1} denoted $\nabla_W \mathcal{H}(\mu_{n+1})$ such that:*

$$\mathcal{G}(\mu_{n+1}) \leq \mathcal{G}(\mu_n) - \gamma \left(1 - \frac{L\gamma}{2}\right) \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2,$$

where we use the notation $\nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})$ to denote $\nabla_W \mathcal{H}(\mu_{n+1}) \circ X_{n+1}$.

Hence, the sequence $(\mathcal{G}(\mu_n))_n$ is decreasing as soon as the step-size is small enough.

4.3 Discrete EVI

To prove a discrete EVI and obtain convergence rates, we need the additional convexity assumption **A2** on the potential function F. We firstly prove the two following lemmas.

Lemma 4.5 *Assume* $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$, $\gamma < 1/L$. *Then for* $n \geq 0$ *and* $\pi \in \mathcal{P}_2(\mathcal{X})$, *there exists a strong Fréchet subgradient at* μ_{n+1} *denoted* $\nabla_W \mathcal{H}(\mu_{n+1})$ *such that:*

$$W^2(\mu_{n+1}, \pi) \leq W^2(\nu_{n+1}, \pi) - 2\gamma \left(\mathcal{H}(\mu_{n+1}) - \mathcal{H}(\pi)\right) - \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})\|_{\mu_{n+1}}^2.$$

Lemma 4.6 *Assume* $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$ *and* $\gamma \leq 1/L$, **A1** *and* **A2** *with* $\lambda \geq 0$. *Then for* $n \geq 0$, *and* $\pi \in \mathcal{P}_2(\mathcal{X})$

$$W^2(\nu_{n+1}, \pi) \leq (1 - \gamma\lambda)W^2(\mu_n, \pi) - 2\gamma \left(\mathcal{E}_F(\mu_n) - \mathcal{E}_F(\pi)\right) + \gamma^2 \|\nabla F\|_{\mu_n}^2.$$

We can now provide a discrete EVI for the functional $\mathcal{G} = \mathcal{E}_F + \mathcal{H}$.

Proposition 4.7 (discrete EVI) *Assume* $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$, $\gamma < 1/L$, **A1** *and* **A2** *with* $\lambda \geq 0$. *Then for* $n \geq 0$ *and* $\pi \in \mathcal{P}_2(\mathcal{X})$, *there exists a strong Fréchet subgradient at* μ_{n+1} *denoted* $\nabla_W \mathcal{H}(\mu_{n+1})$ *such that the Forward Backward Euler scheme verifies:*

$$W^2(\mu_{n+1}, \pi) \leq (1 - \gamma\lambda)W^2(\mu_n, \pi) - 2\gamma \left(\mathcal{G}(\mu_{n+1}) - \mathcal{G}(\pi)\right). \tag{24}$$

4.4 Convergence rates

When the potential function F is convex, we easily get rates from the discrete EVI inequality provided above. Theorem 4.8 is a direct consequence of Proposition 4.7 by taking $\pi = \mu_\star$, and its corollaries provide rates depending on the strong convexity parameter of F.

Theorem 4.8 *Assume* $\mu_0 \in \mathcal{P}_2^r(\mathcal{X})$, $\gamma < 1/L$, **A1** *and* **A2** *with* $\lambda \geq 0$. *Then for every* $n \geq 0$,
$$W^2(\mu_{n+1}, \mu_\star) \leq (1 - \gamma\lambda)W^2(\mu_n, \mu_\star) - 2\gamma \, \mathrm{KL}(\mu_{n+1}|\mu_\star).$$

Corollary 4.9 (Convex case rate) *Under the assumptions of Theorem 4.8, for* $n \geq 0$:

$$\mathrm{KL}(\mu_n|\mu_\star) \leq \frac{W^2(\mu_0, \mu_\star)}{2\gamma n}.$$

Corollary 4.10 (Strongly convex case rate) *Under the assumptions of Theorem 4.8, if* $\lambda > 0$, *then for* $n \geq 0$:

$$W^2(\mu_n, \mu_\star) \leq (1 - \gamma\lambda)^n W^2(\mu_0, \mu_\star).$$

Hence, as soon as F is convex, we get sublinear rates in terms of the Kullback-Leibler divergence, while when F is λ-strongly convex with $\lambda > 0$, we get linear rates in the Wasserstein distance for the iterates of the Forward Backward Euler scheme.

Remark 2 *The rates match the rates of the proximal gradient algorithm in the convex and strongly convex cases by replacing the squared Euclidean norm by the squared Wasserstein distance and the objective function by the KL divergence, [24]. Moreover, these rates are discrete time analogues of the continuous time rates obtained in [1, Th. 11.2.1] for the gradient flow of $\mathcal{G}$.*

The Forward Backward Euler scheme thus enjoys strong theoretical guarantees. However, the implementation of the JKO operator of the entropy functional remains a technical challenge. In the next section, we propose a practical implementation of this operator for discrete measures.

5   JKO operator of the entropy functional

The most demanding step of the FB scheme is the backward step (21). As explained above, the implementation of this step has been investigated in several papers. In this section, we derive an approximate implementation inspired by [13]. First, note that

$$\min_{\nu \in \mathcal{P}_2(\mathcal{X})} \gamma \mathcal{H}(\nu) + \frac{1}{2} W^2(\nu, \mu) = \min_{\nu \in \mathcal{P}_2(\mathcal{X})} \min_{\upsilon \in \Gamma(\mu,\nu)} \gamma \mathcal{H}(\nu) + \frac{1}{2} \int \|x - y\|^2 d\upsilon(x, y)$$

$$= \min_{\upsilon : P_{\#}\upsilon = \mu} \gamma \mathcal{H}(Q_{\#}\upsilon) + \frac{1}{2} \int \|x - y\|^2 d\upsilon(x, y), \quad (25)$$

where P and Q are defined in Section 2.2. Consider the case where the FB scheme is approximated by discretizing the space $\mathcal{X}$ (by a grid for example). In this case, all probability distributions are approximated by discrete measures (supported by the nodes of the grid). From (25), we can obtain a discrete formulation of the JKO of the entropy as follows. Suppose that μ and ν are discrete, *i.e.*, of the form $\mu = \sum_{i=1}^n \mu(i)\delta_{x(i)}$ and $\nu = \sum_{i=1}^m \nu(j)\delta_{y(j)}$ where the coefficients $(\mu(i))$ are known but not the $(\nu(j))$. Then, a discrete reformulation of (25) is

$$\min_{P \in \mathbb{R}_+^{n \times m}} \gamma H(P^T 1_n) + \frac{1}{2}\langle P, M \rangle \text{ such that } P1_m = \mu. \quad (26)$$

where $M \in \mathbb{R}_+^{n \times m}$ is the cost matrix $M(i, j) = \|x(i) - y(j)\|^2$ and $H(a) = \sum_{i=1}^m a(i)\log(a(i))$ for $a = (a(1), \dots, a(m)) \in \mathbb{R}_+^M$. Moreover, $\nu = P^T 1_n$ is the discretized $\text{JKO}_{\gamma\mathcal{H}}(\mu)$, where P is the solution to (26). The discrete distribution ν can be computed in closed form. Introduce the Lagrangian multiplier α associated to the constraint of (26). The Lagrangian reads:

$$\mathscr{L}(P, \alpha) = \langle P, M \rangle + \gamma \sum_{j=1}^n (P^T 1_n)(j) \log((P^T 1_n)(j)) + \langle \alpha, P1_m - \mu \rangle. \quad (27)$$

Now fix $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. Noting that $(P^T 1_n)(j) = \sum_{i=1}^m P(i, j)$, the first order optimality condition results in

$$\frac{\partial \mathscr{L}}{\partial P(i, j)}(P, \alpha) = M(i, j) + \gamma \left(\log((P^T 1_n)(j)) + \frac{P(i, j)}{(P^T 1_n)(j)} \right) + \alpha(i) = 0. \quad (28)$$

Summing Equation (28) for all indices $i \in \{1, \ldots, m\}$, we get successively

$$0 = M^T 1_n(j) + \gamma \left(m \log((P^T 1_n)(j)) + 1 \right) + \langle \alpha, 1_m \rangle, \tag{29}$$

and

$$\nu(j) = (P^T 1_n)(j) = \exp\left(-\frac{M^T 1_n(j) + \langle \alpha, 1_m \rangle + \gamma}{\gamma m} \right). \tag{30}$$

Let $C = \exp\left(-\frac{\langle \alpha, 1_m \rangle + \gamma}{\gamma m} \right)$. Since C does not depend on j,

$$\nu(j) = C \exp\left(-\frac{M^T 1_n(j)}{\gamma m} \right). \tag{31}$$

Moreover as ν is a probability measure, $\sum_{j=1}^m \nu(j) = 1$. Therefore, C can be rewritten as

$$C = \left(\sum_{j=1}^n \exp\left(-\frac{M^T 1_n(j)}{\gamma m} \right) \right)^{-1}. \tag{32}$$

Finally, the discretized $\mathrm{JKO}_{\gamma \mathcal{H}}(\mu)$ is the probability measure $\nu = \sum_{i=1}^m \nu(j) \delta_{y(j)}$ such that

$$\nu \propto \exp\left(-\frac{M^T 1_n(j)}{\gamma m} \right). \tag{33}$$

6   Conclusion

We provided a non-asymptotic analysis of a Forward Backward Euler scheme to sample from $\mu_\star \propto \exp(-F)$. The FB scheme can be seen as a proximal gradient algorithm in the Wasserstein space and enjoys the same convergence rates. Therefore, in terms of number of iterations, the FB scheme might be faster that its alternatives, like Langevin algorithm. However, the FB scheme is challenging to implement in practice, and the iteration complexity of the FB scheme is higher than the iteration complexity of Langevin algorithm. Indeed, the implementation of the FB scheme relies on the computation of the JKO operator of the entropy in the Wasserstein space. Therefore, we studied the JKO operator of the entropy between discrete measures and proved a closed form formula.

The entropy is the regularizer term in Problem (1) for every sampling problem formulated as $\mu_\star \propto \exp(-F)$. Therefore, we stress that the JKO operator of the entropy deserves more investigations. Note that our non-asymptotic analysis of the FB scheme could have been performed for any function $\mathcal{H}$ convex along generalized geodesics and satisfying similar conditions than the entropy, see [1, Eq. 10.1.1a, 10.1.1b]. Finally, in Euclidean spaces, proximal methods go beyond the proximal gradient algorithm, and include for example accelerated, stochastic, variance reduced and primal dual algorithms [5, 12, 17, 29]. The analogues of these methods in the Wasserstein space should lead to improvements and extensions of the FB scheme.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[2] Luigi Ambrosio, Giuseppe Savaré, and Lorenzo Zambotti. Existence and stability for fokker–planck equations with log-concave reference measure. *Probability theory and related fields*, 145(3-4):517–564, 2009.

[3] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[4] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] Espen Bernton. Langevin monte carlo and jko splitting. *arXiv preprint arXiv:1802.08671*, 2018.

[7] Pascal Bianchi, Walid Hachem, and Adil Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *Journal of Convex Analysis*, 2019.

[8] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc., 2014.

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

[10] Haïm Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[11] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for manifold optimization. *arXiv preprint arXiv:1811.00980*, 5(6):8, 2018.

[12] Laurent Condat, Daichi Kitahara, Andrés Contreras, and Akira Hirabayashi. Proximal splitting algorithms: Overrelax them all! *arXiv preprint arXiv:1912.00137*, 2019.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[14] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[15] Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[16] Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559*, 5, 2016.

[17] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

[18] Wen Huang and Ke Wei. Riemannian proximal gradient methods. *arXiv preprint arXiv:1909.06065*, 2019.

[19] 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.

[20] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

[21] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

[22] Bernard Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4(R3):154–158, 1970.

[23] Robert J McCann. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.

[24] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[25] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[26] Juan Peypouquet and Sylvain Sorin. Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. *arXiv preprint arXiv:0905.1270*, 2009.

[27] Adil Salim, Dmitry Kovalev, and Peter Richtárik. Stochastic proximal langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems*, pages 6649–6661, 2019.

[28] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[29] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

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

[31] Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. *arXiv preprint arXiv:1802.08089*, 2018.

A Proof of Lemma 4.1

The map $I - \gamma \nabla F$ is a pushforward from μ to ν. Moreover, denoting $u : x \mapsto \frac{1}{2}\|x\|^2 - \gamma F(x)$, $\nabla u = I - \gamma \nabla F$.

By elementary algebra, for any $(x, y) \in \mathcal{X}^2$ we have:

$$\frac{1}{2}\|x\|^2 = \frac{1}{2}\|y\|^2 - \langle x, y - x \rangle - \frac{1}{2}\|x - y\|^2, \tag{34}$$

and from the smoothness of F,

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \tag{35}$$

Therefore, combining (34) and (35) multiplied by γ gives:

$$\frac{1}{2}\|x\|^2 - \gamma F(x) \leq \frac{1}{2}\|y\|^2 - \gamma F(y) - \langle x - \gamma \nabla F(x), y - x \rangle - \frac{1}{2}(1 - L\gamma)\|x - y\|^2. \tag{36}$$

In other words,

$$u(x) \leq u(y) - \langle \nabla u(x), y - x \rangle - \frac{1}{2}(1 - L\gamma)\|x - y\|^2. \tag{37}$$

Therefore, if $\gamma \leq 1/L$, then u is convex and $\nabla u = T_\mu^\nu$ using Brenier's theorem. Moreover, if $\gamma < 1/L$ then u is $(1 - L\gamma)$-strongly convex. In consequence,

$$(1 - L\gamma)\|x - y\|^2 \leq \langle x - y, \nabla u(x) - \nabla u(y) \rangle.$$

Therefore, ∇u is injective. Furthermore, using the strong convexity of u and [1, Lemma 5.5.3] (see also [1, Th. 6.2.3, Th. 6.2.7]), $\nu \in \mathcal{P}_2^r(\mathcal{X})$.

B Proof of Lemma 4.2

Using [2, Prop. 4.2], $\text{JKO}_{\gamma\mathcal{H}}(\nu)$ is a singleton $\{\mu\}$. Since $D(\mathcal{H}) \subset \mathcal{P}_2^r(\mathcal{X})$, [1, Lemma 10.1.2] implies $\mu \in D(\mathcal{H})$ and $\frac{1}{\gamma}(T_\mu^\nu - I) \in \partial\mathcal{H}(\nu)$ is a strong subgradient of $\mathcal{H}$ at ν.

C Proof of Lemma 4.3

Since $\xi \in \partial H(\mu)$, for every $\phi \in L^2(\mu)$,

$$\mathcal{H}((I + \varepsilon\phi)_{\#}\mu) \geq \mathcal{H}(\mu) + \varepsilon\langle \xi, \phi \rangle_\mu + o(\varepsilon).$$

Applying the last inequality to $\phi = T_\nu^\pi \circ T_\mu^\nu - I$ and using the transfer lemma ($\mu = T_{\nu \#}^\mu \nu$) we have

$$\mathcal{H}((T_\nu^\mu + \varepsilon(T_\nu^\pi - T_\nu^\mu))_{\#}\nu) = \mathcal{H}((I + \varepsilon\phi)_{\#}\mu),$$

$$\langle \xi, \phi \rangle_\mu = \langle \xi(T_\nu^\mu), T_\nu^\pi - T_\nu^\mu \rangle_\nu,$$

and

$$\frac{\mathcal{H}((T_\nu^\mu + \varepsilon(T_\nu^\pi - T_\nu^\mu))_{\#}\nu) - \mathcal{H}(\mu)}{\varepsilon} \geq \langle \xi(T_\nu^\mu), T_\nu^\pi - T_\nu^\mu \rangle_\nu + o(1). \qquad (38)$$

Using the generalized geodesic convexity of the entropy (see [1, Chap. 9] and Equation (12)),

$$\mathcal{H}((T_\nu^\mu + \varepsilon(T_\nu^\pi - T_\nu^\mu))_{\#}\nu) \leq \varepsilon \mathcal{H}(\pi) + (1 - \varepsilon)\mathcal{H}(\mu).$$

Plugging the last inequality into (38),

$$\mathcal{H}(\pi) - \mathcal{H}(\mu) \geq \langle \xi(T_\nu^\mu), T_\nu^\pi - T_\nu^\mu \rangle_\nu + o(1). \qquad (39)$$

We get the conclusion by letting $\varepsilon \to 0$.

D   Proof of Theorem 4.4

Denote $Y_{n+1} := T_{\nu_{n+1}}^{\mu_{n+1}}$ the optimal transport map between ν_{n+1} and μ_{n+1} and $\nabla_W \mathcal{H}(\mu_{n+1})$ the strong Fréchet subgradient of the entropy evaluated at μ_{n+1} defined by Lemma 4.2: $T_{\mu_{n+1}}^{\nu_{n+1}} = I + \gamma \nabla_W \mathcal{H}(\mu_{n+1})$. Since $\mu_{n+1}, \nu_{n+1} \in \mathcal{P}_2^r(\mathcal{X})$, $(I + \gamma \nabla_W H(\mu_{n+1})) \circ Y_{n+1} = I$ using Brenier's theorem. We thus have ν_{n+1}-a.e.:

$$Y_{n+1} = I - \gamma \nabla_W H(\mu_{n+1})(Y_{n+1}). \qquad (40)$$

We firstly bound the entropy term. By taking $\mu = \mu_{n+1}, \pi = \mu_n$ and $\nu = \nu_{n+1}$ in Lemma 4.3, we have:

$$\mathcal{H}(\mu_{n+1}) \leq \mathcal{H}(\mu_n) - \langle \nabla_W \mathcal{H}(\mu_{n+1})(T_{\nu_{n+1}}^{\mu_{n+1}}), T_{\nu_{n+1}}^{\mu_n} - T_{\nu_{n+1}}^{\mu_{n+1}} \rangle_{\nu_{n+1}}. \qquad (41)$$

We now identify $T_{\nu_{n+1}}^{\mu_n}$ and $T_{\nu_{n+1}}^{\mu_{n+1}}$. Recall that $Y_{n+1} = T_{\nu_{n+1}}^{\mu_{n+1}}$. Moreover, using Brenier's theorem and Lemma 4.1, $\nu_{n+1} \in \mathcal{P}_2^r(\mathcal{X})$ and $T_{\nu_{n+1}}^{\mu_n} = (I - \gamma \nabla F)^{-1}$. Therefore,

$$\mathcal{H}(\mu_{n+1}) \leq \mathcal{H}(\mu_n) - \langle \nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1}), (I - \gamma \nabla F)^{-1} - Y_{n+1} \rangle_{\nu_{n+1}}.$$

Using the transfer lemma, with $Y_{n+1} = X_{n+1} \circ (I - \gamma \nabla F)^{-1}$, the last inequality is equivalent to

$$\mathcal{H}(\mu_{n+1}) \leq \mathcal{H}(\mu_n) - \langle \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}), I - X_{n+1} \rangle_{\mu_n}. \qquad (42)$$

Then, we can bound the potential term. Using Equation (40), and $X_{n+1} = Y_{n+1} \circ (I - \gamma \nabla F)$, we have

$$X_{n+1} = I - \gamma \nabla F - \gamma \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}). \qquad (43)$$

Since F is L-smooth, we have [4],

$$F(z) \leq F(x) + \langle \nabla F(x), z - x \rangle + \frac{L}{2}\|x - z\|^2, \qquad \forall\, x, z \in \mathcal{X}. \qquad (44)$$

17

Replacing z by $X_{n+1}(x)$, we obtain

$$F(X_{n+1}(x)) \leq F(x) - \gamma \langle \nabla F(x), \nabla F(x) + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}(x)) \rangle \qquad (45)$$

$$+ \frac{L\gamma^2}{2} \|\nabla F(x) + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}(x))\|^2. \qquad (46)$$

Integrating w.r.t. μ_n,

$$\mathcal{E}_F(\mu_{n+1}) \leq \mathcal{E}_F(\mu_n) - \gamma \langle \nabla F, \nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}) \rangle_{\mu_n}$$

$$+ \frac{L\gamma^2}{2} \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2. \qquad (47)$$

Then, recalling (43) and summing equations (42) and (47), we get

$$\mathcal{H}(\mu_{n+1}) + \mathcal{E}_F(\mu_{n+1}) \leq \mathcal{H}(\mu_n) + \mathcal{E}_F(\mu_n)$$

$$- \gamma \langle \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}), \nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}) \rangle_{\mu_n}$$

$$- \gamma \langle \nabla F, \nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}) \rangle_{\mu_n} + \frac{L\gamma^2}{2} \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2$$

$$\leq \mathcal{H}(\mu_n) + \mathcal{E}_F(\mu_n) - \gamma \left(1 - \frac{L\gamma}{2} \right) \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2.$$

E   Proof of Lemma 4.5

Recall (40),
$$Y_{n+1} = I - \gamma \nabla_W H(\mu_{n+1})(Y_{n+1}) = T_{\nu_{n+1}}^{\mu_{n+1}}.$$
Since $(Y_{n+1}, T_{\nu_{n+1}}^\pi)_{\#} \nu_{n+1}$ is a coupling between μ_{n+1} and π, we can upper bound the Wasserstein distance between μ_{n+1} and π as:

$$W^2(\mu_{n+1}, \pi) \leq \|Y_{n+1} - T_{\nu_{n+1}}^\pi\|_{\nu_{n+1}}^2$$

$$= \|I - T_{\nu_{n+1}}^\pi\|_{\nu_{n+1}}^2 - 2\gamma \langle \nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1}), I - T_{\nu_{n+1}}^\pi \rangle_{\nu_{n+1}} + \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1})\|_{\nu_{n+1}}^2$$

$$= \|I - T_{\nu_{n+1}}^\pi\|_{\nu_{n+1}}^2 - 2\gamma \langle \nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1}), Y_{n+1} - T_{\nu_{n+1}}^\pi \rangle_{\nu_{n+1}} - \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1})\|_{\nu_{n+1}}^2.$$
$$(48)$$

where $\|I - T_{\nu_{n+1}}^\pi\|_{\nu_{n+1}}^2 = W^2(\nu_{n+1}, \pi)$. Moreover, using Lemma 4.3 with $\mu = \mu_{n+1}$ and $\nu = \nu_{n+1}$,

$$-2\gamma \langle \nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1}), Y_{n+1} - T_{\nu_{n+1}}^\pi \rangle_{\nu_{n+1}} \leq -2\gamma \left(\mathcal{H}(\mu_{n+1}) - \mathcal{H}(\pi) \right).$$

Plugging the latter inequality into (48), we get the result.

F   Proof of Lemma 4.6

Since $(I - \gamma \nabla F, T_{\mu_n}^\pi)_{\#} \mu_n$ is a coupling between ν_{n+1} and π, we can upper bound the Wasserstein distance between ν_{n+1} and π as:

$$W^2(\nu_{n+1}, \pi) \leq \|(I - \gamma \nabla F) - T_{\mu_n}^\pi\|_{\mu_n}^2$$

$$= \|I - T_{\mu_n}^\pi\|_{\mu_n}^2 - 2\gamma \langle \nabla F, I - T_{\mu_n}^\pi \rangle_{\mu_n} + \gamma^2 \|\nabla F\|_{\mu_n}^2. \qquad (49)$$

where $\|I - T_{\mu_n}^\pi\|_{\mu_n}^2 = W^2(\mu_n, \pi)$. Moreover, since F is λ-strongly convex, we have:

$$F(x) \le F(z) + \langle \nabla F(x), x - z \rangle - \frac{\lambda}{2} \|x - z\|^2, \qquad \forall\, x, z \in \mathcal{X}. \tag{50}$$

Replacing z by $T_{\mu_n}^\pi(x)$ and multiplying by 2γ, we obtain

$$-2\gamma \langle \nabla F(x), x - T_{\mu_n}^\pi(x) \rangle \le -2\gamma \left(F(x) - F \circ T_{\mu_n}^\pi(x) \right) - \gamma\lambda \|x - T_{\mu_n}^\pi(x)\|^2.$$

Integrating w.r.t. μ_n results in

$$-2\gamma \langle \nabla F, I - T_{\mu_n}^\pi \rangle_{\mu_n} \le -2\gamma \left(\mathcal{E}_F(\mu_n) - \mathcal{E}_F(\pi) \right) - \gamma\lambda W^2(\mu_n, \pi).$$

Plugging the latter inequality into (49) gives the result.

G   Proof of Proposition 4.7

Recall that $Y_{n+1} = T_{\nu_{n+1}}^{\mu_{n+1}}$. Combining Lemma 4.5 and Lemma 4.6, we firstly get

$$\begin{aligned} W^2(\mu_{n+1}, \pi) \le &(1 - \gamma\lambda)W^2(\mu_n, \pi) - 2\gamma \left(\mathcal{E}_F(\mu_n) + \mathcal{H}(\mu_{n+1}) - \mathcal{E}_F(\pi) - \mathcal{H}(\pi) \right) \\ &+ \gamma^2 \|\nabla F\|_{\mu_n}^2 - \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1})\|_{\nu_{n+1}}^2. \end{aligned} \tag{51}$$

Multiplying (47) by 2γ,

$$\begin{aligned} -2\gamma \mathcal{E}_F(\mu_n) \le &- 2\gamma \mathcal{E}_F(\mu_{n+1}) \\ &- 2\gamma^2 \langle \nabla F, \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}) \rangle_{\mu_n} - 2\gamma^2 \|\nabla F\|_{\mu_n} + L\gamma^3 \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2. \end{aligned}$$

Moreover, using the transfer lemma, $\|\nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2 = \|\nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1})\|_{\nu_{n+1}}^2$.
Therefore,

$$\begin{aligned} &-2\gamma \mathcal{E}_F(\mu_n) + \gamma^2 \|\nabla F\|_{\mu_n}^2 - \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})(Y_{n+1})\|_{\nu_{n+1}}^2 \\ &\le - 2\gamma \mathcal{E}_F(\mu_{n+1}) - \gamma^2 \|\nabla F\|_{\mu_n}^2 - \gamma^2 \|\nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2 - 2\gamma^2 \langle \nabla F, \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1}) \rangle_{\mu_n} \\ &\quad + L\gamma^3 \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2 \\ &\le - 2\gamma \mathcal{E}_F(\mu_{n+1}) - \gamma^2 (1 - L\gamma) \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2. \end{aligned}$$

Plugging the last inequality into (51),

$$\begin{aligned} W^2(\mu_{n+1}, \pi) \le &(1 - \gamma\lambda)W^2(\mu_n, \pi) - 2\gamma \left(\mathcal{E}_F(\mu_{n+1}) + \mathcal{H}(\mu_{n+1}) - \mathcal{E}_F(\pi) - \mathcal{H}(\pi) \right) \\ &- \gamma^2 (1 - \gamma L) \|\nabla F + \nabla_W \mathcal{H}(\mu_{n+1})(X_{n+1})\|_{\mu_n}^2. \end{aligned}$$

H   Proof of Theorem 4.8

By using Lemma 3.1, $\mathrm{KL}(\mu_{n+1}|\mu_\star) = \mathcal{G}(\mu_{n+1}) - \mathcal{G}(\pi)$. Plugging into Proposition 4.7 gives the result.

I   Proof of Corollary 4.9

Let $\mathcal{L}_n := 2\gamma n \operatorname{KL}(\mu_n|\mu_\star) + W^2(\mu_n, \mu_\star)$. From Lemma 3.1 and Theorem 4.4, $\operatorname{KL}(\mu_{n+1}|\mu_\star) \leq \operatorname{KL}(\mu_n|\mu_\star)$ if $\gamma < 1/L$. Therefore,

$$2\gamma n \operatorname{KL}(\mu_{n+1}|\mu_\star) + 2\gamma \operatorname{KL}(\mu_{n+1}|\mu_\star) + W^2(\mu_{n+1}, \mu_\star) \leq 2\gamma n \operatorname{KL}(\mu_n|\mu_\star) + W^2(\mu_n, \mu_\star),$$

where we used Theorem 4.8 with $\lambda = 0$ (recall that $\lambda \geq 0$). In other words, $\mathcal{L}_{n+1} \leq \mathcal{L}_n$. Finally,

$$2\gamma n \operatorname{KL}(\mu_n|\mu_\star) \leq \mathcal{L}_n \leq \mathcal{L}_0 = W^2(\mu_0, \mu_\star).$$

J   Proof of Corollary 4.10

Since the KL divergence is nonnegative, from Theorem 4.8,

$$W^2(\mu_{n+1}, \mu_\star) \leq (1 - \gamma\lambda) W^2(\mu_n, \mu_\star).$$

We get the result by iterating.