Journal of Machine Learning Research 25 (2024) 1-66 Submitted 7/23; Revised 5/24; Published 8/24

Wasserstein Proximal Coordinate Gradient Algorithms

Rentian Yao RENTIAN2@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana–Champaign
Champaign, IL 61820, USA

Xiaohui Chen XIAOHUIC@USC.EDU
Department of Mathematics
University of Southern California
Los Angeles, CA 90089, USA

Yun Yang YY84@UMD.EDU
Department of Mathematics
University of Maryland
College Park, MD 20742, USA

Editor: Matthew Hoffman

Abstract

Motivated by approximation Bayesian computation using mean-field variational approximation and the computation of equilibrium in multi-species systems with cross-interaction, this paper investigates the composite geodesically convex optimization problem over multiple distributions. The objective functional under consideration is composed of a convex potential energy on a product of Wasserstein spaces and a sum of convex self-interaction and internal energies associated with each distribution. To efficiently solve this problem, we introduce the Wasserstein Proximal Coordinate Gradient (WPCG) algorithms with parallel, sequential, and random update schemes. Under a quadratic growth (QG) condition that is weaker than the usual strong convexity requirement on the objective functional, we show that WPCG converges exponentially fast to the unique global optimum. In the absence of the QG condition, WPCG is still demonstrated to converge to the global optimal solution, albeit at a slower polynomial rate. Numerical results for both motivating examples are consistent with our theoretical findings.

Keywords: composite convex optimization, coordinate descent, optimal transport, Wasserstein gradient flow, mean-field variational inference, multi-species systems.

1 Introduction

The task of minimizing a functional over the space of probability distributions is common in statistics and machine learning, with a wide range of applications in nonparametric statistics (Kiefer and Wolfowitz, 1956; Yan et al., 2023), Bayesian analysis (Ghosh et al., 2022; Lambert et al., 2022; Trillos and Sanz-Alonso, 2020; Yao and Yang, 2022), online learning (Ballu and Berthet, 2022; Guo et al., 2022), single cell analysis (Lavenant et al., 2021), and spatial economies (Blanchet et al., 2016). Since a probability distribution is an infinite-dimensional object with rich geometric structures, analysis of such an optimization problem requires special treatment and usually leverages optimal transport theory where the

convergence is characterized through the Wasserstein metric. In this paper, we consider the problem of jointly minimizing an objective functional over m distributions $\{\rho_j\}_{j=1}^m$, where $\rho_j \in \mathscr{P}_2(\mathcal{X}_j)$, the space of all probability distributions over the j-th (Euclidean) domain $\mathcal{X}_j$ with finite second-order moments. Precisely, we consider the following general optimization problem:

$$\min_{\rho_1,\dots,\rho_m} \mathcal{F}(\rho_1,\dots,\rho_m) := \mathcal{V}(\rho_1,\dots,\rho_m) + \sum_{j=1}^m \mathcal{H}_j(\rho_j) + \sum_{j=1}^m \mathcal{W}_j(\rho_j), \tag{1}$$

$$\text{where} \qquad \mathcal{V}(\rho_1,\dots,\rho_m) = \int_{\mathcal{X}} V(x_1,\dots,x_m)\,\mathrm{d}\rho_1\cdots\mathrm{d}\rho_m, \tag{2a}$$

$$\mathcal{H}_j(\rho_j) = \int_{\mathcal{X}_j} h_j(\rho_j(x_j))\,\mathrm{d}x_j,\ \forall\, j \in [m], \tag{2b}$$

$$\mathcal{W}_j(\rho_j) = \iint_{\mathcal{X}_j\times\mathcal{X}_j} W_j(x_j,x_j')\,\mathrm{d}\rho_j(x_j)\mathrm{d}\rho_j(x_j'). \tag{2c}$$

Here we use $\mathcal{X} = \bigotimes_{j=1}^m \mathcal{X}_j$ to denote the product space of $\{\mathcal{X}_j\}_{j=1}^m$. Throughout the paper, the notation for a distribution, such as ρ, may stand for both the corresponding probability measure or its density function relative to the Lebesgue measure. In this formulation, $\mathcal{V}$ can be interpreted as an *interaction* potential energy functional with potential function $V : \mathcal{X} \to \mathbb{R}$, describing the interactions among the m input distributions, $\mathcal{H}_j$ is the individual *internal* energy functional associated with ρ_j for some convex function $h_j : [0,\infty) \to \mathbb{R}$, and $\mathcal{W}_j$ is the *self-interaction* functional associated with ρ_j for some self-interaction potential function $W_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$. By convention, we define $\mathcal{H}_j(\rho_j) = \infty$ if ρ_j is not absolutely continuous with respect to the Lebesgue measure and h_j is not a constant function. When $h_j \equiv 0$ and $W_j \equiv 0$, this problem degenerates into the problem of minimizing an m-variate function V on the Euclidean space; when $h_j(x) = x\log x$, $W_j = 0$, and $m = 1$, this optimization problem amounts to minimize the Kullback-Leibler (KL) divergence $\mathrm{KL}(\cdot \,\|\, \rho_1^*)$ with $\rho_1^* \propto e^{-V}$. In general, an internal energy functional $\mathcal{H}_j$ with h_j satisfying $h_j(x) \to \infty$ as $x \to \infty$ prevents the optimal solution of ρ_j from degenerating into a point mass measure. Our problem of minimizing the functional in (1) is chiefly motivated by the two representative examples below.

Example 1 (**Mean-field inference in variational Bayes**). In approximate Bayesian computation (ABC) when optimizing a functional over a single high-dimensional distribution to approximate the posterior distribution, it can often be computationally and theoretically advantageous to employ a mean-field approximation by breaking into the product of many lower-dimensional ones, provided that the bias introduced by this approximation is tolerable or can be suitably controlled. Specifically, consider a Bayesian model with a likelihood function $p(x \,|\, \theta)$ and a prior density function π (relative to the Lebesgue measure) over the parameter space $\Theta = \bigotimes_{j=1}^m \Theta_j$, where the parameter $\theta = (\theta_1,\dots,\theta_m)$ is divided into m pre-specified blocks with $\theta_j \in \Theta_j$. Given observed i.i.d. data $X^n = (X_1,\dots,X_n)$, the posterior density function of θ can be expressed via Bayes' rule as

$$\pi_n(\theta) := p(\theta \,|\, X^n) = \frac{\pi(\theta)\prod_{i=1}^n p(X_i \,|\, \theta)}{\int_\Theta \pi(\theta)\prod_{i=1}^n p(X_i \,|\, \theta)\,\mathrm{d}\theta}. \tag{3}$$

We use Π_n to denote the corresponding posterior distribution. In practice, π_n is often computationally intractable due to the lack of an explicit form of the normalizing constant, i.e., the denominator in (3). Mean-field variational inference (MFVI) (Bishop and Nasrabadi, 2006), which approaches this task by turning the integration problem into an optimization problem, can be formulated as finding a closest fully factorized distribution $\widehat{\pi}_n = \bigotimes_{j=1}^m \widehat{\rho}_j$ to approximate the target posterior distribution Π_n with respect to the KL divergence; that is, computing

$$\widehat{\pi}_n = \operatorname*{argmin}_{\rho=\bigotimes_{j=1}^m \rho_j} \mathrm{KL}(\rho \,\|\, \Pi_n), \quad \text{s.t.} \quad \rho_j \in \mathscr{P}(\Theta_j) \quad \forall\, j \in [m].$$

Up to a ρ independent constant, this is equivalent to

$$\widehat{\pi}_n = \operatorname*{argmin}_{\rho=\bigotimes_{j=1}^m \rho_j} \left\{ \int_\Theta \big[nU_n(\theta) - \log \pi(\theta) \big] \,\mathrm{d}\rho_1 \cdots \mathrm{d}\rho_m + \sum_{j=1}^m \int_{\Theta_j} \rho_j \log \rho_j \right\}, \tag{4}$$

where $U_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i \,|\, \theta)$ corresponds to the (averaged) negative log-likelihood function of data X^n. It is then apparent that above is a special case of the general formulation (1) by taking $V = nU_n - \log \pi$, $h_j(x) = x \log x$, $W_j = 0$, and $\mathcal{X}_j = \Theta_j$. The optimal solution $(\widehat{\rho}_1, \ldots, \widehat{\rho}_m)$ corresponds to the mean-field approximation of the joint posterior distribution of the parameter $\theta = (\theta_1, \ldots, \theta_m)$ via the relationship $\widehat{\pi}_n = \bigotimes_{j=1}^m \widehat{\rho}_j$. Ghosh et al. (2022) shows the connection between MFVI (4) and the objective functional (1) without proposing practical algorithm for solving (4). Yao and Yang (2022) focuses on solving special cases of problem (4) with $m = 2$ blocks corresponding to continuous model parameters and discrete latent variables.

Concurrent and independent to our work, Arnese and Lacker (2024) and Lavenant and Zanella (2024) have also investigated the minimization of (4) with coordinate ascent variational inference (CAVI) using optimal transport theory. Arnese and Lacker (2024) explored the convergence of CAVI with parallel and sequential update schemes, while Lavenant and Zanella (2024) studied the random update scheme. Both papers primarily emphasized the theoretical analysis of the CAVI algorithm, demonstrating similar convergence results to ours for CAVI when the posterior distributions are log-concave. In contrast, our work introduces a novel algorithm that differs from CAVI and studies its algorithmic convergence rate. Furthermore, our analysis is conducted under the quadratic growth condition (detailed in Assumption D), which is weaker than the log-concavity assumption as later shown in Proposition 8. □

Example 2 (**Equilibrium in multi-species systems with cross-interaction)**. Multi-species systems (Carrillo et al., 2018; Daus et al., 2022) arise in applications in cell biology (Pinar, 2021) and population dynamics (Zamponi and Jüngel, 2017). In this example, we consider the following non-local multi-species cross-interaction model with diffusion,

$$\partial_t \rho_j = \nabla \cdot \Big(\rho_j \Big[\nabla V_j - \sum_{i=1}^m (\nabla K_{ij}) * \rho_i + \nabla h_j'(\rho_j) \Big] \Big), \quad j \in [m], \tag{5}$$

where $\rho_j(x,t)$ is the unknown mass density of species j at location x and time t, V_j is the external potential field affecting species j, K_{jj} is the self-interaction potential of species j,

and $\{K_{ij} : 1 \leqslant i \neq j \leqslant m\}$ are the cross-interaction potentials. When $\mathcal{H}_j$ is the negative self-entropy functional $h_j(\rho_j) = \rho_j \log \rho_j$ for the j-th species, PDE (5) corresponds to the Fokker–Planck equation associated with the mean-field multi-species stochastic interacting particle system (Daus et al., 2022),

$$\mathrm{d}X_j(t) = -\nabla V_j\big(X_j(t)\big)\,\mathrm{d}t + \sum_{i=1}^{m} \big(\nabla K_{ij} * \rho_i(\cdot,t)\big)\big(X_j(t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_j(t), \quad \forall\, j \in [m]. \quad (6)$$

In particular, $\rho_j(\cdot,t)$ corresponds to the density function associated with the distribution of $X_j(t)$ with initialization $X_j(0) \sim \rho_j(\cdot,0)$ for all $j \in [m]$. Another common class of $\mathcal{H}_j$ has $h_j(\rho_j) = \rho_j^{m_j}$ with $m_j > 1$, corresponding to diffusion in porous media (Aronson, 2006; Vázquez, 2007). For such an entropy functional $\mathcal{H}_j$, there is no analogous stochastic differential equation (SDE) corresponding to equation (5). For symmetric multi-species models where $K_{ij} = K_{ji}$, finding the equilibrium of equation (5) is equivalent to solving the optimization problem (1) with $V(x) = \sum_{i=1}^m V_i(x_i) - \sum_{1\leqslant i<j\leqslant m} K_{ij}(x_i - x_j)$ and $W_j(x_j, x_j') = -K_{jj}(x_j - x_j')/2$ (cf. Proposition 25). The symmetric interaction kernel assumption is natural in physics applications due to Newton's third law of motion.

The demand of computing the stationary distribution (or equilibrium) of a multi-species system naturally arises in physical science, such as chemical engineering (Carrayrou et al., 2002; Paz-Garcia et al., 2013). However, most existing literature considers computational methods for calculating the equilibrium of interacting particle systems with only one species. For example, Gutleb et al. (2022) consider the case where $m = 1$, $V = h = 0$, and the interaction potential W has attractive-repulsive power-law form, by approximating the equilibrium measure with a series of orthogonal polynomials. Under their approximating schemes, the original problem of finding the equilibrium turns into an optimization problem of solving the coefficient of each polynomial. In the multi-species setting, Owolabi and Atangana (2019) studied the equilibrium of multi-species fractional reaction-diffusion systems by simply approximating the spatial derivatives with second-order Taylor expansion without analyzing the approximation error of their methods. □

1.1 Our Contributions

In this paper, we propose the *Wasserstein Proximal Coordinate Gradient* (WPCG) algorithm as a tool to minimize a composite geodesically convex functional of form (1) over *multiple* distributions, which extends the coordinate descent algorithm from the Euclidean space to the space of probability distributions. We provide detailed convergence analysis for WPCG with three updating schemes: *parallel*, *sequential*, and *random*. Specifically, we show that: (i) under the condition that the potential function V is smooth and convex, and $\{h_j\}_{j=1}^m$ and $\{W_j\}_{j=1}^m$ are convex, WPCG converges to a global optimal solution of (1) at rate $O(1/k)$, where k denotes the iteration count; (ii) under the additional quadratic growth (QG) condition in (17), WPCG converges to the unique global optimum exponentially fast in the Wasserstein-2 metric. Note that QG condition is a weaker requirement than the strong convexity on V and/or $\{h_j\}_{j=1}^m$ and $\{W_j\}_{j=1}^m$; see Figure 1 for the implications of various assumptions for proving convergence in optimization and Section 3 for further details. In addition, implementation of WPCG with the parallel updating scheme inherently

supports parallelization, which enhances its computational scalability to high-dimensional problems.

Previous studies, including (Ambrosio et al., 2005; Salim et al., 2020; Wibisono, 2018), predominantly concentrate on using Wasserstein gradient flow for the minimization of a functional over a single distribution. To the best of our knowledge, the current work is among the first study to:

1. investigate the application of Wasserstein gradient flow for minimizing a functional over multiple distributions;
2. extend the design and analysis of coordinate descent type algorithms for optimization in Euclidean spaces to those for optimization in Wasserstein spaces.

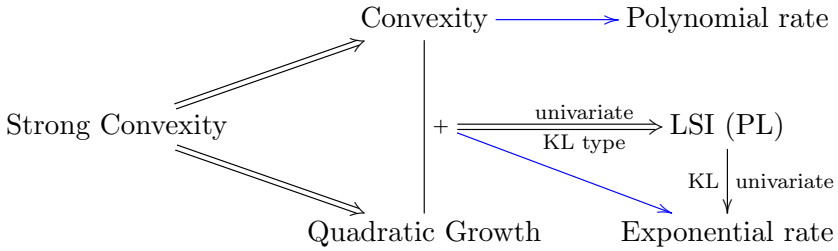


Figure 1: Implication diagram of commonly used assumptions for proving convergence results in optimization. Double arrows show the connections between assumptions on the Wasserstein space: strong convexity implies convexity and the quadratic growth condition; convexity and the quadratic growth condition together imply LSI for KL-type functionals. Single arrows show the convergence rates implied by different assumptions, among which the blue ones are studied in this work for minimizing multivariate objective functionals.

We emphasize two key differences from optimization on the Euclidean space.

First, most common internal energy functionals $\{\mathcal{H}_j\}_{j=1}^m$ in (2) (e.g., negative self-entropy) are **non-smooth** in the usual sense—the difference incurred by the linearization of $\mathcal{H}_j$ around any ρ cannot be upper bounded by a multiple of the Wasserstein distance to ρ. For coordinate descent optimization on the Euclidean space, such a smoothness property serves as two critical purposes: (i) it ensures that the discretization error of gradient descent is compensated by the contraction of the corresponding continuous-time gradient flow; (ii) it prevents overshooting of the coordinate descent update by controlling the misalignment between the coordinate-wise steepest descent direction and the entire steepest descent direction, so that coordinate descent algorithm keeps making progress in minimizing the objective functional. In the Wasserstein space, (explicit) gradient/coordinate ascent algorithms with non-smooth $\mathcal{H}_j$ require the linearization of $\mathcal{H}_j$ in the algorithm and a uniform control of smoothness of the Wasserstein gradient of $\mathcal{H}_j$. Our analysis reveals that considering an implicit scheme (or a proximal type update) in designing coordinate descent type algorithms in Wasserstein spaces can avoid imposing a smoothness assumption on $\{\mathcal{H}_j\}_{j=1}^m$ when analyzing their convergence. In particular, the linear convergence of WPCG only

requires the potential gradient ∇V to be Lipschitz (cf. Assumption B in Section 3). Thus, convergence rates of WPCG are derived under more realistic conditions on the objective function $\mathcal{F}$ and hold for general functionals $\{\mathcal{H}_j\}_{j=1}^m$ than the negative self-entropy. This allows us to compute the stationary measure of non-local multi-species degenerate cross diffusion beyond heat diffusion (cf. more details for our numeric example on porous medium equation in Section 5.2).

A second challenge in demonstrating the convergence of WPCG lies in the difficulty of proving a Polyak–Łojasiewicz (PL) type inequality, which requires some positive power of certain norm of the "Wasserstein gradient" of target functional $\mathcal{F}$ at any $\rho_{1:m} := (\rho_1, \dots, \rho_m)$ to be lower bounded by a multiple of $\mathcal{F}(\rho_{1:m}) - \mathcal{F}(\rho^*_{1:m})$, where $\rho^*_{1:m}$ denotes a global minimizer of $\mathcal{F}$. It is known (Karimi et al., 2016) that proving a PL inequality is crucial and common for analyzing gradient-based optimization algorithms in the Euclidean case. A PL inequality is typically implied by strong convexity or a log-Sobolev inequality (LSI). It is also known that a quadratic growth (QG) condition with convexity is strictly weaker than the strong convexity (see discussions after Assumption D). While the QG condition together with convexity implies LSI for the KL-type functional over a single distribution, there is no corresponding result for a general functional over multiple distributions $\rho_{1:m}$. The main technical difficulty comes from the lack of tensorization for a "multivariate" Wasserstein gradient.

1.2 Related Work

Optimization over one distribution. The seminal paper by Jordan, Kinderlehrer, and Otto (Jordan et al., 1998) introduces an iterative scheme (now commonly known as the *JKO scheme*) serving as the time-discretization of a continuous dynamic in the space of probability distributions following the direction of the steepest descent of a target functional with respect to the Wasserstein-2 metric. In later developments, the JKO scheme has been recognized as a promising numerical method for optimizing a functional over a probability distribution and can be viewed as a special proximal gradient update (Rockafellar, 1997) relative to the Wasserstein-2 metric in the space of probability distributions (see Section 2.2 for more details). Therefore, in the remainder of the paper, we will primarily use the term Wasserstein proximal gradient scheme instead of the JKO scheme, as this terminology more clearly suggests its close link with general convex optimization.

Recently, several other discretization methods are proposed for minimizing a functional over a single distribution (i.e., $m = 1$ in our setting). Different from the Wasserstein proximal gradient scheme that can be viewed as an implicit (backward) scheme for discretizing the Wasserstein gradient flow (WGF), some explicit (forward) schemes, analogous to the usual gradient descent on the Euclidean space, are considered and analyzed for solving specific problems. For example, Chewi et al. (2020) analyze a gradient descent algorithm for computing the barycenter on the Bures–Wasserstein manifold of centered Gaussian probability measures. One key ingredient in their convergence analysis is to prove a quadratic growth condition for the barycenter functional that leads to a PL inequality and implies the exponential convergence of the algorithm. For the KL divergence functional $\mathrm{KL}(\,\cdot\,\|\,\rho^*)$ with $\rho^* \propto e^{-V}$ for some strongly convex and smooth potential function V, Wibisono (2018) shows that a symmetrized Langevin algorithm, which is an implementable discretization of the

Langevin dynamics, reduces the bias in the classical unadjusted Langevin algorithm (Durmus and Moulines, 2017) and attains exponential convergence up to a step size dependent discretization error. Salim et al. (2020) consider a more general setting where the objective functional consists of the same potential functional and a general non-smooth term that is convex along generalized geodesics on the Wasserstein space; they propose a forward-backward algorithm, which runs a forward step (gradient descent) for the potential functional and a backward step (proximal gradient) for the non-smooth term, and prove its exponential convergence under the same conditions on potential V.

Optimization over special cases of multiple distributions. Several recent works focus on solving an important special case of problem (1) in the context of MFVI as we described in Example 1, where a multi-dimensional Bayesian posterior distribution is approximated by the product of several lower-dimensional distributions relative to the KL divergence. Yao and Yang (2022) consider a block MFVI for Bayesian latent variable models with two blocks, one for the discrete latent variables and one for the continuous model parameters, which corresponds to problem (1) with $m = 2$. They propose and analyze a majorization-minimization algorithm for solving MFVI, which can be viewed as a distributional extension of the classical expectation–maximization (EM) algorithm. Due to the special property that minimizing the latent variable block in the problem admits a closed-form expression, their algorithm is effectively a time-discretized WGF for optimizing a single distribution, with an effective potential function V changing over the iterations. When the population-level log-likelihood function is locally strictly concave, they show that their algorithm converges exponentially fast to the solution of MFVI. Ghosh et al. (2022) study a WGF-based algorithm for solving MFVI without latent variables. Their study shows that the discretized flow (Wasserstein proximal gradient scheme) converges to a mathematically well-defined continuous flow when the step size is small, assuming certain conditions on the potential function V. However, their analysis is asymptotic and they do not directly examine the convergence of the time-discretized algorithm, leaving it unclear if the discrete dynamic system is stable so that the discretization error does not accumulate over time and whether an explicit convergence rate can be obtained.

The rest of this paper is organized as follows. In Section 2, we present an overview of essential concepts in optimal transport that are necessary for developing our methods and theory. Section 3 introduces the WPCG algorithm with three different update schemes: parallel, sequential, and random update schemes. We also propose two different numerical methods to solve the Wasserstein proximal gradient scheme, a key step for implementing the WPCG algorithms. Theoretical analyses of the WPCG algorithms are provided in Section 4 in the presence and absence of a λ-quadratic growth condition. In Section 5, we demonstrate the applications of our algorithm and theory to mean-field variational inference and multi-species systems, along with numerical experiments. In Section 6, we conclude the paper and discuss some open problems for potential future research.

1.3 Notations

For any $\mathcal{X} \subset \mathbb{R}^d$, let $\mathscr{P}(\mathcal{X})$ be the collection of all probability measures on $\mathcal{X}$, $\mathscr{P}_2(\mathcal{X}) = \{\mu : \mathbb{E}_{X\sim\mu}[\|X\|^2] < \infty\}$ be the set of probability measures with finite second-order moments, and $\mathscr{P}_2^r(\mathcal{X}) \subset \mathscr{P}_2(\mathcal{X})$ be the subset that includes all absolutely continuous probability

measures with respect to the Lebesgue measure on $\mathcal{X}$. For a measurable map $T : \mathcal{X} \to \mathcal{X}$, let $T_{\#} : \mathscr{P}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ be its corresponding pushforward operator, i.e., $\nu = T_{\#}\mu$ if and only if $\nu(A) = \mu(T^{-1}(A))$ for any measurable set $A \subset \mathcal{X}$. Throughout the paper, we assume $\mathcal{X}$ to be convex and compact unless otherwise stated.

Let $\pi^0(x, y) = x$ and $\pi^1(x, y) = y$ be two projection maps (projecting into the first and second components). Let $\Pi(\mu, \nu) = \{\gamma : (\pi^0)_{\#}\gamma = \mu, (\pi^1)_{\#}\gamma = \nu\}$ be the set of couplings[1], or all transport plans, between μ and ν. Let $\Pi_o(\mu, \nu)$ be the set of all optimal transport plans, whose precise definition can be found in Section 2 below. For a random variable X, we use $\mathcal{L}(X)$ to denote its distribution.

Let $C^1(\mathcal{X})$ be the set of functions on $\mathcal{X}$ that have continuous derivatives. Let $L^1(\mathcal{X})$ and $L^\infty(\mathcal{X})$ represent the class of integrable functions and uniformly bounded functions on $\mathcal{X}$, respectively. For a vector $x = (x_1, \ldots, x_m)$, we assume its sub-vector without the j-th entry is denoted by the shorthand $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m)$. Let $\|\cdot\|$ denote the vector ℓ_2 norm. For a matrix A, let $|||A|||_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$ denote its matrix $\ell_2 \to \ell_2$ operator norm.

2 Preliminaries on Optimal Transport

In this section, we briefly review some concepts and results in optimal transport that are necessary for explaining and developing our methods and analysis.

2.1 Wasserstein Space and Geodesics

For any $\mu, \nu \in \mathscr{P}_2(\mathcal{X})$, the Wasserstein-2 distance between μ and ν is defined as

$$\mathrm{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^2 \, \mathrm{d}\pi(x, y) \right\}, \tag{7}$$

It is known that $(\mathscr{P}_2(\mathcal{X}), \mathrm{W}_2)$ is a metric space called Wasserstein space. Moreover, if $\mathcal{X}$ is complete and separable, $(\mathscr{P}_2(\mathcal{X}), \mathrm{W}_2)$ is complete as well (Bolley, 2008). The infimum in (7) always admits a solution γ called an *optimal transport plan* (Santambrogio, 2015). When $\mu \in \mathscr{P}_2^r(\mathcal{X})$, the optimal transport plan is unique and takes form $\gamma = (\mathrm{Id}, T_\mu^\nu)_{\#}\mu$, implying $\nu = (T_\mu^\nu)_{\#}\mu$, where Id is the identity map and T_μ^ν is called the *optimal transport map* from μ to ν (Santambrogio, 2015). Moreover, Brenier's Theorem (Brenier, 1987) states that this unique optimal transport map $T_\mu^\nu = \nabla\psi$ is the gradient of a convex function ψ.

Note that $(\mathscr{P}_2(\mathcal{X}), \mathrm{W}_2)$ is not a flat metric space, but is positively curved in the Alexandrov sense (Ambrosio et al., 2005). Geodesics on a curved space are the shortest paths connecting two points. $(\mathscr{P}_2(\mathcal{X}), \mathrm{W}_2)$ is indeed a geodesic space. For any $\mu_0, \mu_1 \in \mathscr{P}_2^r(\mathcal{X})$ and $\gamma \in \Pi_o(\mu_0, \mu_1)$, the constant-speed geodesic connecting μ_0 and μ_1 is $\mu_t = (\pi_t)_{\#}\gamma$, where $\pi_t = (1 - t)\pi^0 + t\pi^1$.

1. A coupling between two distributions μ and ν is a joint distribution whose two marginals are μ and ν.

2.2 Convex Functionals on Wasserstein Space

Convexity is an important concept in optimization. On the Euclidean space, a function $f : \mathcal{X} \to \mathbb{R}$ is λ-*strongly convex* if

$$f(x_t) \leqslant (1-t)f(x_0) + tf(x_1) - \frac{t(1-t)\lambda}{2}\|x_1 - x_0\|^2,$$

and we say f is L-*smooth* if

$$f(x_t) \geqslant (1-t)f(x_0) + tf(x_1) - \frac{t(1-t)L}{2}\|x_1 - x_0\|^2,$$

for all $t \in [0,1]$ where $x_t = (1-t)x_0 + tx_1$. In convex optimization, L-smoothness and λ-strong convexity are common sufficient conditions to guarantee the exponential convergence rate of an (explicit) gradient descent algorithm, $x^{k+1}_{\text{grad}} = x^k_{\text{grad}} - \tau\nabla f(x^k_{\text{grad}})$, $k = 0, 1, \ldots$, towards the unique global minimum of f. In comparison, the following proximal gradient algorithm

$$x^{k+1}_{\text{prox}} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f(x) + \frac{1}{2\tau}\|x - x^k_{\text{prox}}\|^2 \right\}, \quad k = 0, 1, \ldots, \tag{8}$$

only requires the λ-strong convexity to guarantee its exponential convergence rate (Beck, 2017). Note that a proximal gradient algorithm is also called an *implicit* gradient descent algorithm, since the first-order optimality condition (FOC) for (8) reads $x^{k+1}_{\text{prox}} = x^k_{\text{prox}} - \tau\nabla f(x^{k+1}_{\text{prox}})$. The theoretical advantage of the proximal scheme comes at the price of the additional computational cost of numerically solving this FOC equation.

Similar to these notions in the Euclidean case, a functional $\mathcal{F} : \mathscr{P}_2(\mathcal{X}) \to (-\infty, +\infty]$ to be λ-*strongly convex along geodesics*, if for all $\mu_0, \mu_1 \in \mathscr{P}_2(\mathcal{X})$, we have

$$\mathcal{F}(\mu_t) \leqslant (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\lambda}{2}t(1-t)\,\mathrm{W}_2^2(\mu_0, \mu_1), \tag{9}$$

where $\{\mu_t : 0 \leqslant t \leqslant 1\}$ is the constant speed geodesic curve connecting μ_0 and μ_1. When $\lambda = 0$, we simply say that $\mathcal{F}$ is *geodesically convex*. It can be shown that both functionals $\mathcal{W}_j(\rho_j)$ and $\mathcal{H}_j(\rho_j)$ in our problem formulation (4) are geodesically convex when their defining functions W_j and h_i are convex (Ambrosio et al., 2005, Chapter 9.3). However, different from the Euclidean case where all convex functions are continuous, geodesically convex functionals may not be continuous with respect to the W_2 distance.

Example 3 (Geodesically convex functional may not be continuous). Consider the negative self-entropy functional $\rho \mapsto \int_{\mathcal{X}} \rho\log\rho$ which is geodesically convex. Let ρ_∞ be the uniform distribution on $[0,1]$ and ρ_n be the uniform distribution on A_n where $A_n = \bigcup_{i=0}^{n-1}[\frac{2i}{2n}, \frac{2i+1}{2n}]$. Then $\mathrm{W}_2(\rho_n, \rho_\infty) \to 0$ as $n \to \infty$, while $\int \rho_n \log\rho_n = \log 2$ for all n and $\int \rho_\infty \log\rho_\infty = 0$. □

As we mentioned in the introduction, the Wasserstein proximal gradient scheme, an iterative algorithm defined as below, is commonly used for minimizing a functional $\mathcal{F} : \mathscr{P}_2(\mathcal{X}) \to (-\infty, +\infty]$ defined on the Wasserstein space,

$$\rho^{k+1} = \operatorname*{argmin}_{\rho\in\mathscr{P}_2(\mathcal{X})} \left\{ \mathcal{F}(\rho) + \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho, \rho^k) \right\}, \quad k = 0, 1, \ldots, \tag{10}$$

with the initial distribution ρ^0 and the step size $\tau > 0$. This is analogous to the proximal gradient algorithm (8), i.e., the backward time-discretization of a gradient flow on the Euclidean space. The convergence of Wasserstein proximal gradient scheme for the KL divergence functional $\mathcal{F}_{\mathrm{KL}}(\rho) = \int V \,\mathrm{d}\rho + \int \rho \log \rho$ is well studied in literature. As on the Euclidean space, piecewise interpolation $\rho_t = \rho^{\lfloor \frac{t}{\tau} \rfloor}$ converges to a continuous dynamics on Wasserstein space, i.e., the solution of Fokker–Planck equation $\partial_t \rho_t = \nabla \cdot (\rho_t \nabla (V + \log \rho_t))$ (Santambrogio, 2015). Since the Fokker–Planck equation is also the density evolution equation of Langevin dynamics (Jordan et al., 1998)

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{11}$$

one can also use the Euler–Maruyama method to numerically approximate the Wasserstein proximal gradient scheme for the KL using particle approximation. It is shown in the literature that when V is strongly convex, the solution $\{\rho_t : t \geqslant 0\}$ of the Fokker–Planck equation converges to the unique global minimizer $\rho^* \propto e^{-V}$ of $\mathcal{F}_{\mathrm{KL}}$ (Villani, 2021). Corollary 2.8 in (Yao and Yang, 2022) further shows that the time-discretization of the continuous time dynamic via Wasserstein proximal gradient scheme $\{\rho^k : k \in \mathbb{Z}_+\}$ is unbiased, and also converges to $\rho^* \propto e^{-V}$ exponentially fast when V is strongly convex.

In this paper, we are primarily interested in functional $\mathcal{F} : \mathscr{P}_2(\mathcal{X}_1) \times \cdots \times \mathscr{P}_2(\mathcal{X}_m) \to (-\infty, \infty]$ that is defined over m distributions. Analogously, we say that $\mathcal{F}$ is *(blockwise) λ-strongly convex* if

$$\mathcal{F}(\mu_1^t, \ldots, \mu_m^t) \leqslant (1-t)\mathcal{F}(\mu_1^0, \ldots, \mu_m^0) + t\mathcal{F}(\mu_1^1, \ldots, \mu_m^1) - \frac{\lambda}{2} t(1-t) \sum_{j=1}^m \mathrm{W}_2^2(\mu_j^0, \mu_j^1)$$

for all $\mu_j^0, \mu_j^1 \in \mathscr{P}_2(\mathcal{X}_j)$ and the corresponding constant speed geodesics $\{\mu_j^t : 0 \leqslant t \leqslant 1\}$, $\forall j \in [m]$. In the next section, we will extend the aforementioned Wasserstein proximal gradient scheme from minimizing a simple functional of one distribution into minimizing a functional of multiple distributions by describing several coordinate ascent versions of JKO and analyze their convergence.

2.3 First Variation and Wasserstein Gradient

Let $\mathcal{F} : \mathscr{P}_2(\mathcal{X}) \to (-\infty, +\infty]$ be a lower semi-continuous functional. For a measure $\mu \in \mathscr{P}_2^r(\mathcal{X})$, we call μ to be *regular* for $\mathcal{F}$ if $\mathcal{F}((1-\varepsilon)\mu + \varepsilon\nu) < \infty$ for every $\varepsilon \in [0,1]$ and every absolutely continuous probability measure $\nu \in \mathscr{P}(\mathcal{X}) \bigcap L^\infty(\mathcal{X})$. If μ is regular for $\mathcal{F}$, the *first variation* (or Gateaux derivative) of $\mathcal{F}$ at μ is a map $\frac{\delta \mathcal{F}}{\delta \rho}(\mu) : \mathcal{X} \to \mathbb{R}$ satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{F}(\mu + \varepsilon\chi)\bigg|_{\varepsilon=0} = \int_{\mathcal{X}} \frac{\delta \mathcal{F}}{\delta \rho}(\mu)\,\mathrm{d}\chi$$

for any perturbation $\chi = \tilde{\mu} - \mu$ such that $\tilde{\mu} \in \mathscr{P}_2^r(\mathcal{X})$ and $\int \mathrm{d}\chi = 0$. Clearly, $\frac{\delta \mathcal{F}}{\delta \rho}(\mu)$ is uniquely defined up to additive constant. This is analogous to the gradient of $\mathcal{F}$ in the $L^2(\mathcal{X})$ sense. If $\mathcal{F}$ is a geodesically convex functional defined on the Wasserstein space, and $\mathcal{F}(\mu) < \infty$ at some μ, then the vector field $\xi = \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\mu) \in L^2(\mu; \mathcal{X})$ will satisfy (Ambrosio et al., 2005)

$$\mathcal{F}(\nu) \geqslant \mathcal{F}(\mu) + \int_{\mathcal{X}} \left\langle \xi, T_\mu^\nu - \mathrm{Id} \right\rangle \mathrm{d}\mu, \ \forall\, \nu \in \mathscr{P}_2^r(\mathcal{X}). \tag{12}$$

We say that $\xi = \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\mu)$ is a *strong subdifferential* (or Fréchet derivative) of $\mathcal{F}$ at μ. Strong subdifferential captures the "gradient" with respect to the W_2 metric and is more convenient to use than the first variation when analyzing the convergence of a gradient flow on the Wasserstein space. Henceforth, we will also refer to a strong subdifferential of $\mathcal{F}$ as its Wasserstein gradient.

3 Wasserstein Proximal Coordinate Gradient (WPCG) Algorithms

In this section, we introduce the WPCG algorithm with three different yet common update schemes: parallel update (WPCG-P), sequential update (WPCG-S), and random update (WPCG-R).

3.1 Parallel Update

In the WPCG-P algorithm, each coordinate is updated synchronously, i.e., we solve the following Wasserstein proximal gradient scheme (13) for all $j \in [m]$ at the same time to get ρ_j^k,

$$\rho_j^{k+1} = \underset{\rho_j \in \mathscr{P}_2^r(\mathcal{X}_j)}{\operatorname{argmin}} \left\{ \mathcal{V}(\rho_j, \rho_{-j}^k) + \mathcal{H}_j(\rho_j) + \mathcal{W}_j(\rho_j) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho_j, \rho_j^k) \right\}. \tag{13}$$

Recall that we have used the shorthand $\rho_{-j}^k = (\rho_1^k, \ldots, \rho_{j-1}^k, \rho_{j+1}^k, \ldots, \rho_m^k)$ to denote the collection of all distributions at iteration k except for the j-th one, for each $k \in \mathbb{N}$ and $j \in [m]$. When $\mathcal{H}_j = \mathcal{W}_j = 0$, the solution ρ_j^{k+1} of the subproblem (13) satisfies $(\mathrm{Id} + \nabla V_j^k)_{\#} \rho_j^{k+1} = \rho_j^k$, where $V_j^k = \int V(x_j, x_{-j}) \, \mathrm{d}\rho_{-j}^k$ is a function on $\mathcal{X}_j$. Furthermore, when the initialization ρ_j^0 is a point mass for all $j \in [m]$, the subproblem degenerates to the coordinate proximal descent method for updating the j-th block. The parallel update scheme can be parallelly computed and therefore is preferred when the number of coordinates m is large. However, as we shall see in Section 4, this scheme is more sensitive to the step size compared with the other two (i.e., sequential and random) update schemes since it may diverge when the step size τ is large. Pseudo-code of WPCG-P is shown in Algorithm 1.

Algorithm 1 WPCG-P

Initialize distribution $\rho^0 = (\rho_1^0, \ldots, \rho_m^0)$ arbitrarily, number of iterations T, and step size τ.
for $k = 0$ **to** $T - 1$ **do**
  for $j = 1$ **to** m **do**
    $\rho_j^{k+1} = \operatorname{argmin}_{\rho_j \in \mathscr{P}_2^r(\mathcal{X}_j)} \mathcal{V}(\rho_j, \rho_{-j}^k) + \mathcal{H}_j(\rho_j) + \mathcal{W}_j(\rho_j) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho_j, \rho_j^k)$;
  end for
end for

3.2 Sequential Update

In the WPCG-S algorithm, the updates are made sequentially through all coordinates, one by one, at each iteration. Although WPCG-S cannot be made parallel, the functional

value is always convergent regardless of the step size τ magnitude since WPCG-S is always a descent algorithm. To ease the presentation of the result, we use the notation $\rho^k_{i:j} = (\rho^k_i, \ldots, \rho^k_j)$ to denote the distributions at iteration k from index i to j, with the convention that $\rho^k_{i:j} = \varnothing$ if $i > j$. Pseudo-code of WPCG-S is shown in Algorithm 2.

Algorithm 2 WPCG-S

Initialize distribution $\rho^0 = (\rho^0_1, \ldots, \rho^0_m)$ arbitrarily, number of iterations T, and step size τ.
for $k = 0$ **to** $T - 1$ **do**
  for $j = 1$ **to** m **do**
    $\rho^{k+1}_j = \operatorname{argmin}_{\rho_j \in \mathscr{P}_2^r(\mathcal{X}_j)} \mathcal{V}(\rho^{k+1}_{1:(j-1)}, \rho_j, \rho^k_{(j+1):m}) + \mathcal{H}_j(\rho_j) + \mathcal{W}_j(\rho_j) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho_j, \rho^k_j)$;
  end for
end for

3.3 Random Update

In the WPCG-R algorithm, we sample the index j_l of the coordinate to be updated randomly and uniformly from $[m]$, independently of the previous selections of indices. To make the convergence rate comparable to the other two schemes, we consider one iteration of WPCG-R as the process of updating M randomly selected coordinates, where the batch size M is of the order $O(m \log m)$. This ensures that with high probability, each coordinate has been updated at least once per iteration. Similar to WPCG-S, WPCG-R is also a descent algorithm regardless of the choice of the step size τ. Pseudo-code of WPCG-R is shown in Algorithm 3.

Algorithm 3 WPCG-R

Initialize distribution $\rho^0 = (\rho^0_1, \ldots, \rho^0_m)$ arbitrarily, number of iterations T, step size τ, and the number M of coordinates (or batch size) updated in each iteration.
for $k = 0$ **to** $T - 1$ **do**
  $\rho^{k,0} = \rho^k$;
  for $l = 0$ **to** $M - 1$ **do**
    Choose index $j_l \sim \mathrm{unif}([m])$;
    $\rho^{k,l+1}_{j_l} = \operatorname{argmin}_{\rho_{j_l} \in \mathscr{P}_2^r(\mathcal{X}_{j_l})} \mathcal{V}(\rho_{j_l}, \rho^{k,l}_{-j_l}) + \mathcal{H}_{j_l}(\rho_{j_l}) + \mathcal{W}_{j_l}(\rho_{j_l}) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho_{j_l}, \rho^{k,l}_{j_l})$;
    $\rho^{k,l+1}_{-j_l} = \rho^{k,l}_{-j_l}$;
  end for
  $\rho^{k+1} = \rho^{k,M}$;
end for

3.4 Implementation

Note that a common key step in the WPCG algorithm is to solve the Wasserstein proximal gradient scheme (10) with $\mathcal{F}(\rho) = \mathcal{V}(\rho) + \mathcal{H}(\rho) + \mathcal{W}(\rho)$, which does not have an explicit

solution. Here we introduce two numerical methods: function approximation (FA) approach and particle approximation via SDE.

3.4.1 Particle Approximation via SDE

The first method is particle approximation via SDE, which is only applicable when $\mathcal{H}(\rho)$ is the negative self-entropy functional. Recall that the Wasserstein proximal gradient scheme is an implicit scheme for discretizing the WGF. When $\mathcal{V}(\rho) = \int V \,\mathrm{d}\rho$, $\mathcal{H}(\rho) = \int \rho \log \rho$, and $\mathcal{W}(\rho) = \iint W(x, x') \,\mathrm{d}\rho(x)\mathrm{d}\rho(x')$, the WGF of $\mathcal{F}$ starting from ρ_0 is the evolution of the distribution of the following SDE,

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t - \Big(\int \nabla_1 W(X_t, x) + \nabla_2 W(x, X_t)\,\mathrm{d}\rho_t(x)\Big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad X_t \sim \rho_t, \tag{14}$$

where ∇_1 and ∇_2 are the gradients with respect to the first and the second variates of W. This connection between SDE and WGF motivates us to discretize the WGF by discretizing its corresponding SDE. When the step size (for discretization) is small, these two different discretization schemes are expected to be close. Then, we can approximate the Wasserstein proximal gradient scheme by the evolution of the discretized SDE through the empirical measure of particles which satisfy the following updating formula

$$X_b^{k+1} - X_b^k = -\Big(\nabla V(X_b^k) + \frac{1}{B}\sum_{b'=1}^{B}\big[\nabla_1 W(X_b^k, X_{b'}^k) + \nabla_2 W(X_{b'}^k, X_b^k)\big]\Big)\tau + \sqrt{2\tau}\eta_b^k,$$

where τ is the step size, B is the number of particles, and $\eta_b^k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for all $k \in \mathbb{Z}_+$ and $b \in [B]$. The last recursive equation is the discretized representation of the dynamics (14) for approximating the Wasserstein proximal gradient scheme (10), and ρ^{k+1} can be approximated by the empirical distribution of $\{X_b^{k+1} : b \in [B]\}$.

3.4.2 Function Approximation Method

The SDE approach is no longer applicable for a more general internal energy functional $\mathcal{H}$ beyond the negative self-entropy. Instead, we leverage the function approximation methods converting Wasserstein proximal gradient scheme (10) into an optimization problem over the function space. Note that finding ρ^{k+1} that minimizes (10) is equivalent to finding a transport map T such that $T_\#\rho^k$ minimizes (10). Precisely, we have the following statement.

Proposition 4. If $\rho^k \in \mathscr{P}_2^r$ and

$$T^{k+1} = \operatorname*{argmin}_T \ \mathcal{V}(T_\#\rho^k) + \mathcal{H}(T_\#\rho^k) + \mathcal{W}(T_\#\rho^k) + \frac{1}{2\tau}\int \|T(x) - x\|^2 \,\mathrm{d}\rho^k, \tag{15}$$

then $\rho^{k+1} = T_\#^{k+1}\rho^k$ minimizes (10).

We highlight the optimization problem (15) is *unconstrained*. By Brenier's Theorem, the optimal transport map from ρ^k to ρ^{k+1} is the gradient of a convex function. Mokrov et al. (2021) consider the constrained optimization problem by restricting $T = \nabla\psi$ to the

gradient of a convex function ψ and finding the optimal convex function ψ^{k+1} by using input-convex neural networks (ICNN) (Amos et al., 2017). On the contrary, Proposition 4 shows that an unconstrained optimization problem is enough; directly minimizing the objective functional (15) yields the optimal transport map T^{k+1} from ρ^k to ρ^{k+1}. The intuition is that, when T^{k+1} is not the optimal transport map, changing T^{k+1} to the optimal transport map $\widetilde{T}^{k+1}$ from ρ^k to ρ^{k+1} does not change the first three functional values in (15) but decreases the last term, which contradicts to the optimality of T^{k+1}.

In practice, the optimization problem in Proposition 4 over the function space can be solved by function approximation methods such as kernel methods and neural networks. Moreover, since we do not have access to the time-evolving density ρ^k, we still need to approximate the objective functional in Proposition 4. Suppose $\{X_b^k : b \in [B]\}$ are samples drawn from ρ^k. We provide two concrete examples, which will be used later in Section 5, to show how to approximate the objective functional in Proposition 4. The details will be postponed to Appendix C.7.

Example 5 (Approximating negative self-entropy). For $\mathcal{H}(\rho^k) = \int \rho^k \log \rho^k$, we can consider the optimization problem

$$T^{k+1} = \operatorname*{argmin}_T \ \frac{1}{B}\sum_{b=1}^{B} V(T(X_b^k)) - \frac{1}{B}\sum_{b=1}^{B} \log|\det \nabla T(X_b^k)| + \frac{1}{B^2}\sum_{b,b'=1}^{B} W\big(T(X_b^k), T(X_{b'}^k)\big) + \frac{1}{2B\tau}\sum_{b=1}^{B} \|T(X_b^k) - X_b^k\|^2.$$

□

Example 6 (Approximating porous medium type diffusion). For $\mathcal{H}(\rho^k) = \int (\rho^k)^n$, we can consider the following optimization problem

$$T^{k+1} = \operatorname*{argmin}_T \ \frac{1}{B}\sum_{b=1}^{B} V(T(X_b^k)) + \frac{1}{B}\sum_{b=1}^{B} \big[\widehat{\rho}_{\mathrm{kde}}^k(X_b^k) \cdot |\det \nabla T(X_b^k)|^{-1}\big]^{n-1} + \frac{1}{B^2}\sum_{b,b'=1}^{B} W\big(T(X_b^k), T(X_{b'}^k)\big) + \frac{1}{2B\tau}\sum_{b=1}^{B} \|T(X_b^k) - X_b^k\|^2, \tag{16}$$

where $\widehat{\rho}_{\mathrm{kde}}^k$ is the kernel density estimation of ρ^k by $\{X_b^k : b \in [B]\}$. □

Remark 7 (Comparison between FA and SDE approaches). Compared with the SDE approach, the FA approach can be applied when the diffusion term $\mathcal{H}_j$ is beyond the negative self-entropy. According to our numerical results in Section 5, the performance of the FA approach will not be affected when the system contains super-quadratic drift terms. However, smaller step size has to be chosen to apply the SDE approach, since τ needs to be smaller than $O(L_{\mathrm{c}}^{-1})$, where $L_{\mathrm{c}} = \max_{j\in[m],x\in\mathcal{X}} \|\nabla_j^2 V(x)\|_{\mathrm{op}})$ is defined in the discussion after Assumption B (Section 4), so that the explicit discretization of the SDE converges. It then takes more iterations for the SDE approach to converge due to this requirement on a smaller step size. An additional attractive aspect of the FA approach is its unbiasedness for any step size τ. This is due to the fact that any fixed-point solution (from a distributional perspective) to its iterative formula must correspond to a global minimum of $\mathcal{F}$.

In comparison, the SDE approach often exhibits bias—its fixed point usually deviates by $O(\sqrt{\tau})$ from a global minimum of $\mathcal{F}$. Although choosing a decreasing step size sequence when discretizing the SDE (14) can alleviate the constant bias, it does so at the expense of convergence speed. This is evidenced by the slow convergence observed in Figure 7a, which persists even for the largest permissible step size due to the existence of a super-quadratic error term. □

4 Convergence Analysis of WPCG Algorithms

In this section, we derive the convergence rates of the WPCG (-P, -S, -R) algorithms. First, we shall make the following assumptions and discuss their implications. These assumptions are assumed to hold throughout the rest of paper unless otherwise specified.

Assumption A (Internal energy). All $h_1, h_2, \ldots, h_m \in C^1(\mathbb{R}_+)$ are convex and satisfy:

1. For some $\alpha > \frac{d_j}{d_j+2}$, it holds
$$h_j(0) = 0, \quad \liminf_{x_j \downarrow 0_+} \frac{h_j(x_j)}{x_j^\alpha} > -\infty;$$
2. the map $x_j \mapsto x_j^{d_j} h_j(x_j^{-d_j})$ is convex and non-increasing in $\mathbb{R}_+$;
3. $h_j(\rho_j) \in L^1(\mathcal{X}_j)$ implies $\rho_j h_j'(\rho_j) \in L^1(\mathcal{X}_j)$.

The first condition on h_j implies that the negative part $h_j(\rho_j)$ is integrable. Details are referred to Remark 3.9 in (Ambrosio and Savaré, 2007). The second condition implies the convexity of h_j along geodesics (Ambrosio and Savaré, 2007; Santambrogio, 2015). The third condition guarantees that the first variation of $\mathcal{H}_j$ at ρ_j admits the explicit form as $h_j'(\rho_j)$ (Lemma 9 in Appendix). For example, $h_j(x_j) = x_j \log x_j$ (corresponding to negative self-entropy) and $h_j(x_j) = x_j^{m_j}$ with $m_j > 1$ (corresponding to porous medium type diffusion) satisfy all these three conditions.

Assumption B (Potential energy). There exists $L > 0$ such that
$$\|\nabla_j V(x_j, x_{-j}) - \nabla_j V(x_j, x'_{-j})\| \leqslant L\|x_{-j} - x'_{-j}\|, \quad \forall\, j \in [m].$$

This Lipschitz constant is different from the coordinate Lipschitz constant L_{c} used by Wright (2015) in $\|\nabla_j V(y_j, x_{-j}) - \nabla_j V(z_j, x_{-j})\| \leqslant L_{\mathrm{c}}\|y_j - z_j\|$ for all $j \in [m]$. A finite L_{c} guarantees the smoothness of V on each coordinate, which helps control the change of the function value of V after each iteration through the gradient ∇V. Wright (2015) also defined the restricted Lipschitz constant L_{r} as $\|\nabla V(y_j, x_{-j}) - \nabla V(z_j, x_{-j})\| \leqslant L_{\mathrm{r}}\|y_j - z_j\|$ to study the convergence property of asynchronous coordinate GD. We also let L_{g} be the global Lipschitz constant defined through $\|\nabla V(x) - \nabla V(y)\| \leqslant L_{\mathrm{g}}\|x - y\|$, which is commonly used in literature of GD. In fact, we have $0 < L/L_{\mathrm{c}} \leqslant \sqrt{m-1}$, $0 \leqslant L/L_{\mathrm{r}} \leqslant \sqrt{1 - 1/m}$, and $0 \leqslant L/L_{\mathrm{g}} \leqslant 1$. The proof of these connections is provided in Appendix E.

From the above connections, we can see that our Lipschitz assumption is the weakest, which only requires a control on the off-diagonal term of $\nabla^2 V$ (when V is twice differentiable). This difference is due to our choice of the proximal-type optimization algorithm,

where only the magnitude of interaction (off-diagonal) components in V matters in the sense that changing the target function from $V(x) + \sum_{j=1}^m f_j(x_j)$ into V for any univariate functions $\{f_j\}_{j=1}^m$ does not affect the convergence of a proximal coordinate descent algorithm. Technically, this irrelevance of marginal (diagonal) components is due to the reason that convergence of a proximal gradient algorithm for minimizing a univariate function does not require the smoothness (gradient) condition. However, the interaction component, which incurs the overshooting when alternating the coordinate-wise steepest descent direction, cannot be eliminated by using the proximal scheme.

Assumption C (Self-interaction energy). There are functions f_j and g_j such that both

$$\widetilde{V}(x) \coloneqq V(x) - \sum_{j=1}^m \big[f_j(x_j) + g_j(x_j)\big] \quad \text{and} \quad \widetilde{W}_j(x_j, x_j') \coloneqq f_j(x_j) + W_j(x_j, x_j') + g_j(x_j')$$

are convex for all $j \in [m]$.

When W_j is symmetric, we can simply take $f_j = g_j$ for all $j \in [m]$. Without loss of generality, in the remaining of this paper, we will only consider the case where $f_j = g_j = 0$ for all $j \in [m]$. For the general case, note that the functional value $\mathcal{F}(\rho_{1:m})$ in (1) remains the same while replacing V and W_j with $\widetilde{V}$ and $\widetilde{W}_j$, respectively. Thus our proof with $f_j = g_j = 0$ can be easily extended to the general case. This freedom of allocating the marginal components between the potential energy and the self-interaction energy in our analysis is again due to the fact that the convergence of the proximal coordinate algorithm is only affected by the interaction components within V (cf. discussions after Assumption B).

Assumption D (Overall growth). The target functional $\mathcal{F}$ satisfies the following λ-quadratic growth (λ-QG) condition: for $\lambda \geqslant 0$,

$$\mathcal{F}(\rho_{1:m}) - \mathcal{F}(\rho_{1:m}^*) \geqslant \frac{\lambda}{2} \sum_{j=1}^m \mathrm{W}_2^2(\rho_j, \rho_j^*), \quad \forall\, \rho_j \in \mathscr{P}_2^r(\mathcal{X}_j). \tag{17}$$

When $\lambda > 0$, the λ-QG condition guarantees the uniqueness of the solution to the optimization problem (1). When $\lambda = 0$, the minimizer of (1) may not be unique, but the functional value convergences to $\mathcal{F}(\rho_{1:m}^*)$ for any global minimizer $\rho_{1:m}^*$ of (1) (cf. Theorem 18). On the Euclidean space, it is known that λ-strong convexity is stronger than λ-QG condition plus convexity, which together implies a PL inequality (Karimi et al., 2016). A similar result holds in the Wasserstein setting. In fact, the following proposition shows that Assumption D holds when V is λ-strongly convex, which is a sufficient condition for $\mathcal{F}$ being strongly convex. Its proof is provided in Appendix C.

Proposition 8. Assumption D holds when V is λ-strongly convex.

On the other hand, we also want to mention that the λ-QG in Assumption D is strictly weaker than λ-strong convexity. A simple illustrative example is $W_j \equiv 0$, $h_j(\rho_j) = \rho_j \log \rho_j$, and $V(x) = V_1(x_1) + \cdots + V_m(x_m)$ with $V_j(x_j) = \frac{\lambda}{2}\,|x_j|$ for $x_j \in [-1, 1]$ and $V_j(x_j) = \frac{\lambda}{2}\, x_j^2$

elsewhere. In this case, $\mathcal{F}(\rho_{1:m}) = \mathcal{F}_1(\rho_1) + \cdots + \mathcal{F}_m(\rho_m)$ with

$$\mathcal{F}_j(\rho_j) = \int_{\mathcal{X}_j} V_j \,\mathrm{d}\rho_j + \int \rho_j \log \rho_j, \quad \mathcal{X}_j \subset \mathbb{R}.$$

It is straightforward to verify that $\mathcal{F}_j$ satisfies the λ-QG condition due to the growth rate of V_j (see p.280 in (Villani, 2021) for more details) and is convex along geodesics, but is not strongly convex along geodesics since V is not strongly convex in $[-1,1]^m$ (e.g. consider two absolutely continuous measures supported on $[-1,1]^m$).

The PL inequality has an analogy on the Wasserstein space which is called log-Sobolev inequality (LSI) (Villani, 2021). However, it is not easy to directly prove an LSI, mainly due to the following two reasons. Firstly, LSI is often exclusively used to study the KL divergence type objective functionals associated to the relative entropy functional, while we are considering more general objective functionals. Secondly, when we consider a multivariate functional on the Wasserstein space, the multivariate Wasserstein gradient does not tensorize, i.e., the sum of the norm of its blockwise Wasserstein gradient does not equal to the norm of the Wasserstein gradient of the functional treated as a univariate functional (see ahead (18) in Remark 17 below).

4.1 Convergence Rates of WPCG under λ-QG Assumption

Now, we are ready to present our main theoretical results on the convergence rates of the WPCG algorithms. For notation simplicity, we denote $\rho^* := \rho^*_{1:m}$ as a solution of the optimization problem (1) and $\rho^k := \rho^k_{1:m}$ as the solution of the WPCG algorithms (-P, -S, -R) at the k-th iteration when the context has no ambiguity.

Theorem 9 (**Exponential convergence rate of WPCG-P under λ-QG**). Assume Assumptions A, B, C, and D hold for some $\lambda > 0$. Let ρ^k be the solution of WPCG-P in (13) at the k-th iteration. If the step size satisfies $0 < \tau < \frac{m-1/2}{L(m-1)^{3/2}}$, then

$$\mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \frac{2}{\lambda}\big[1 + C_1(m, L, \tau)\lambda\big]^{-k}\big[\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big], \quad \forall\, k \in \mathbb{Z}_+,$$

where

$$C_1(m, L, \tau) = \frac{\frac{1}{2\tau} + (m-1)\big(\frac{1}{\tau} - L\sqrt{m-1}\big)}{2m[2L^2(m-1) + \frac{2}{\tau^2}]} > 0.$$

Remark 10 (**Comments on the step size and iteration complexity of WPCG-P**). (i) By taking $\tau^{-1} = KL\sqrt{m-1}$ for $K > 1$, we have $C_1(m, L, \tau) \geqslant \frac{(K-1)\sqrt{m-1}}{4L(K^2+1)m}$. Therefore, the best iteration complexity based on our theory for achieving ε-accuracy is

$$\frac{\log \frac{2[\mathcal{F}(\rho^0)-\mathcal{F}(\rho^*)]}{\lambda\varepsilon}}{\log(1 + \lambda C_1(m, L, \tau))} \leqslant \frac{\log \frac{2[\mathcal{F}(\rho^0)-\mathcal{F}(\rho^*)]}{\lambda\varepsilon}}{\log(1 + \frac{\lambda(K-1)\sqrt{m-1}}{4L(K^2+1)m})} \leqslant \frac{4L(K^2+1)m}{\lambda(K-1)\sqrt{m-1}} \cdot \frac{\log \frac{2[\mathcal{F}(\rho^0)-\mathcal{F}(\rho^*)]}{\lambda\varepsilon}}{1 - \frac{\lambda(K-1)\sqrt{m-1}}{8L(K^2+1)m}},$$

i.e., WPCG-P requires at most $O(\frac{\sqrt{m}L}{\lambda}\log(\frac{1}{\lambda\varepsilon}))$ iterations to achieve ε-accuracy when the step size τ is on the scale $O(\frac{1}{L\sqrt{m}})$. [In the sequel, we will drop the logarithmic factor in the

iteration complexity when exponential convergence is achieved.] (ii) Both the upper bound of the step size τ and the smoothness condition of V are necessary for the convergence of the parallel update scheme. The following Example 11 shows that the upper bound $O(\frac{1}{L\sqrt{m}})$ of the step size τ cannot be relaxed, i.e., there exists V, $\mathcal{H}_j$, and $\mathcal{W}_j$ such that WPCG-P diverges when τ exceeds the scale $O(\frac{1}{L\sqrt{m}})$. Moreover, when we choose τ on this scale, the iteration complexity is exactly $O(\frac{\sqrt{m}L}{\lambda})$ which coincides with the result in the previous discussion. □

Example 11 (**Optimality of step size and convergence rate of WPCG-P**). Consider the case where $V(x) = \frac{1-\alpha}{2}\|x\|^2 + \frac{\alpha}{2}(x_1 + \cdots + x_m)^2$ and $W_j = h_j = 0$ so that $\rho_j^* = \delta_0$, the point mass measure at 0 for each $j \in [m]$. In this case, when we initialize ρ_j^0 to a point mass measure for all $j \in [m]$, the optimization problem (1) on the Wasserstein space degenerates into an optimization problem on $\mathbb{R}^m$ with objective function V. The (gradient) Lipschitz constant of V in Assumption B is $L = \alpha\sqrt{m-1}$. In each iteration, the update scheme in WPCG-P algorithm is equivalent to solve

$$\begin{aligned} x_j^{k+1} &= \operatorname*{argmin}_{x_j \in \mathbb{R}} \left\{ V(x_j, x_{-j}^k) + \frac{(x_j - x_j^k)^2}{2\tau} \right\} \\ &= \operatorname*{argmin}_{x_j \in \mathbb{R}} \left\{ \frac{1-\alpha}{2} x_j^2 + \frac{\alpha}{2}(x_j + s_{-j}^k)^2 + \frac{(x_j - x_j^k)^2}{2\tau} \right\}, \end{aligned}$$

where we use the shorthand $s^k = x_1^k + \cdots + x_m^k$ to denote the sum of all coordinates at the k-th iterate, and $s_{-j}^k = s^k - x_j^k$. A direct calculation for this example reveals that

$$s^k = \Big(\frac{\tau^{-1} - (m-1)\alpha}{1 + \tau^{-1}}\Big)^k s^0.$$

Therefore, s^k only converges as $k \to \infty$ when $|\tau^{-1} - (m-1)\alpha| \leqslant |1 + \tau^{-1}|$, which is equivalent to

$$\tau \leqslant \frac{2}{(m-1)\alpha - 1} = O\Big(\frac{1}{L\sqrt{m}}\Big).$$

This calculation shows that our upper bound requirement on step size τ is necessary and tight. To study the tightness of the convergence rate implied by the theorem, we note that a direct calculation leads to

$$\begin{pmatrix} x_1^{k+1} \\ \vdots \\ x_m^{k+1} \end{pmatrix} = \frac{1}{1+\tau} \begin{pmatrix} 1 & -\alpha\tau & \cdots & -\alpha\tau \\ -\alpha\tau & 1 & \cdots & -\alpha\tau \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha\tau & -\alpha\tau & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1^k \\ \vdots \\ x_m^k \end{pmatrix} =: A_{\mathrm{p}} x^k.$$

It is easy to verify that the m eigenvalues of the symmetric matrix A_{p} are

$$\lambda_1 = \cdots = \lambda_{m-1} = \frac{1+\alpha\tau}{1+\tau}, \quad \lambda_m = \frac{1 + \alpha\tau - \alpha\tau m}{1+\tau}.$$

Under the optimal step size (modulo constants) $\tau = 1/(L\sqrt{m-1}) = 1/(\alpha(m-1))$, we have

$$\lambda_1 = \cdots = \lambda_{m-1} = \frac{m}{m-1+\frac{1}{\alpha}}, \quad \lambda_m = 0.$$

Therefore, we can always find some initialization $x^0 \in \mathbb{R}^m$ (e.g., any eigenvector associated with λ_1) such that

$$\|x^k\| = \lambda_1^k \|x^0\| = \Big(1 - \frac{1-\alpha}{m\alpha+1-\alpha}\Big)^k \|x^0\|.$$

Under this initialization, the algorithm takes $O(m\alpha/(1-\alpha)) = O(L\sqrt{m}/\lambda)$ iterations to converge. This example shows that our convergence rate bound is unimprovable when τ is $O(\frac{1}{L\sqrt{m}})$. □

Next, we establish the exponential convergence rate of WPCG-S.

Theorem 12 (**Exponential convergence rate of WPCG-S under λ-QG**). Assume Assumptions A, B, C, and D hold for some $\lambda > 0$. Let ρ^k be the solution of WPCG-S at the k-th iteration. Then for any step size $\tau > 0$ we have

$$\mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \frac{2}{\lambda}\big[1 + \lambda C_2(m, L, \tau)\big]^{-k}\Big(\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\Big),$$

where

$$C_2(m, L, \tau) = \Big(8\tau L^2(m-1) + 8\tau^{-1}\Big)^{-1} > 0.$$

Remark 13 (**Comparison with existing results in the Euclidean case**). Similar to the parallel update scheme, by taking $\tau^{-1} = L\sqrt{m-1}$ the iteration complexity also grows at a rate of $O(\frac{L\sqrt{m}}{\lambda})$. When degenerated to optimization on the Euclidean space (by taking $\mathcal{H}_j \equiv 0$ and $\mathcal{W}_j \equiv 0$), our result is comparable to the rate $O(\frac{\sqrt{m}L_{\mathrm{g}}}{\lambda})$ in Theorem 6.3 of (Wright and Recht, 2022) for sequential coordinate gradient descent method with strongly convex and smooth functions. When we take $\tau = L^{-1}$, the iteration complexity turns out to be $O(\frac{Lm}{\lambda})$. This iteration complexity is no larger than $O(\frac{\max_j L_j}{\min_j\{L_j+\mu_j\}} \cdot \frac{mL_{\mathrm{g}}}{\lambda})$ derived by Li et al. (2017) when applying coordinate proximal gradient descent with the sequential update scheme, where λ_j and L_j are the parameters of strong convexity and smoothness of the j-th block, and the step size for updating the j-th block is L_j^{-1}. Their result has an additional factor $\frac{\max_j L_j}{\min_j\{L_j+\mu_j\}}$ because they approximated the smooth part of the objective function by its first-order Taylor expansion before applying the proximal descent step. As we mentioned, we believe that the difference between the Lipschitz constant in our result and the ones chosen in (Li et al., 2017; Wright and Recht, 2022) for optimization problems on the Euclidean space is due to our choice of a proximal-type optimization method. In fact, for the sequential update scheme, the Lipschitz constant can be further relaxed to the lower triangular Lipschitz constant (Hua and Yamashita, 2015). □

Remark 14 (**Effect of step size in WPCG-S**). The iteration complexity $O(\frac{L\sqrt{m}}{\lambda})$ when $\tau^{-1} = L\sqrt{m-1}$ in WPCG-S is smaller than the iteration complexity $O(\frac{mL_{\mathrm{g}}}{\lambda})$ derived by Sun and Ye (2021) for alternating minimization algorithm (corresponding to $\tau = \infty$)

with the sequential update scheme on the Euclidean space, which shows that using an overly large τ can still drag down the convergence speed. It is an interesting open problem that if the iteration complexity $O(\frac{L\sqrt{m}}{\lambda})$ can be improved. Even when it degenerates to an optimization problem on $\mathbb{R}^m$ ($\mathcal{H}_j = \mathcal{W}_j = 0$ and ρ^0 is a point mass measure), the optimality of this rate remains as an open problem. Sun and Ye (2021) showed that there exists an optimization problem (related to the Example 11 by choosing α carefully) on $\mathbb{R}^m$ such that alternating minimization algorithm with the sequential update scheme has iteration complexity $O(\frac{mL_{\mathrm{g}}}{\lambda})$. □

For WPCG-R, we can establish the following rate of convergence.

Theorem 15 (**Exponential convergence rate of WPCG-R under λ-QG**). Let $M = \lceil 2m\log(mL)\rceil$. Assume Assumptions A, B, C, and D hold for some $\lambda > 0$. Let ρ^k be the solution of WPCG-R at the k-th iteration with batch size M. Then for any step size $\tau > 0$, we have

$$\mathbb{E}\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \min\{C_3(m, L, \tau), 1\}^k \big[\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big]$$

where

$$C_3(m, L, \tau) = \left(1 + \frac{\lambda\tau}{8[1 + 2e\tau^2 L^2 m\log(mL)]}\right)^{-1} + \frac{2}{mL^2}.$$

The constant C_3 is a strictly positive constant less than one if

$$\frac{4}{mL^2} < \frac{\lambda}{8[\tau^{-1} + 2e\tau L^2 m\log(mL)]} < 1.$$

Remark 16 (**Comments on the batch size in WPCG-R**). In the proof of Theorem 15, we may control the norm of the blockwise Wasserstein gradient by the distance between two consecutive iterates only when all coordinates have been updated at least once in each iteration. For this reason, we need to choose a larger batch size, $M = \lceil 2m\log(mL)\rceil$, compared to the other two update schemes (parallel and sequential), where precisely $M = m$ coordinates are updated in each iteration. □

Remark 17 (**Comparison with existing results in the Euclidean case**). In optimization problems on the Euclidean space, the random update scheme usually has smaller iteration complexity compared with the other two schemes in worst case scenario, such as in coordinate GD (Wright, 2015) and alternating minimization algorithm (Sun and Ye, 2021). However, on the Wasserstein space, we are only able to derive the same iteration complexity up to a log factor as in the other two update schemes, given $\lambda > 0$ and a fixed step size $\tau > 0$. In particular, when $\tau^{-1} = L\sqrt{2em\log(mL)}$, our result implies an $O(\frac{L\sqrt{m\log(mL)}}{\lambda})$ iteration complexity of WPCG-R, where in each iteration we need to solve $\lceil 2m\log(mL)\rceil$ sub-problems. Recall that there are only m sub-problems in each iteration in WPCG-P and WPCG-S. Therefore, the iteration complexity in WPCG-R is $[\log(mL)]^{3/2}$ times larger than WPCG-P and WPCG-S. This logarithmic factor appears since we need to solve more sub-problems in each iteration compared with the other two update schemes, so that every coordinate has been updated with high probability in each iteration. When $\tau = L^{-1}$, we derive the iteration complexity $O(\frac{mL\log(mL)}{\lambda})$, which is slower than $O(\frac{L_{\mathrm{c}}}{\lambda})$ in the coordinate

GD with random update scheme derived by Wright and Recht (2022) if $mL\log(mL) > L_{\mathrm{c}}$. We conjecture that such a sub-optimal rate when specialized to the Euclidean case (Sun and Ye, 2021; Wright and Recht, 2022) is due to the extra complexity of dealing with the Wasserstein space, where there is no tensorization structure of the Wasserstein gradient. More precisely, the following key identity

$$\mathbb{E}_{j_k}\left\|\nabla_{j_k} f(x^k)\right\|^2 = \frac{1}{m}\sum_{j=1}^{m}\left\|\nabla_j f(x^k)\right\|^2 = \frac{1}{m}\left\|\nabla f(x^k)\right\|^2, \quad \text{where} \quad j_k \sim \mathrm{unif}\left([m]\right)$$

holds in the Euclidean case, indicating on average that randomly selecting a coordinate to update for m times is similar to updating all coordinates at once as in the parallel scheme. So, a majority of the analysis for the parallel update scheme can be reused for analyzing the random update scheme that is expected to have smaller iteration complexity than the other two update schemes for Euclidean optimizations. Nevertheless, the similar identity does not hold for the Wasserstein gradient of potential energy $\mathcal{V}$. Specifically, if we denote $\nabla_{\mathrm{W}_2}\mathcal{V}(\rho) = \nabla V$ and $\nabla_{\mathrm{W}_2,j}\mathcal{V}(\rho) = \int \nabla_j V(x_j, x_{-j})\,\mathrm{d}\rho_{-j}$ as the Wasserstein gradient and the j-th marginal Wasserstein gradient of $\mathcal{V}$ (with respect to ρ_j) at ρ, then $\nabla_{\mathrm{W}_2}\mathcal{V}(\rho)$ does not tensorize, that is,

$$\begin{aligned}\sum_{j=1}^{m}\left\|\nabla_{\mathrm{W}_2,j}\mathcal{V}(\rho)\right\|_{L^2(\rho)}^2 &= \sum_{j=1}^{m}\int_{\mathcal{X}_j}\left\|\int_{\mathcal{X}_{-j}}\nabla_j V(x_j, x_{-j})\,\mathrm{d}\rho_{-j}\right\|^2\mathrm{d}\rho_j \\ &\neq \sum_{j=1}^{m}\int_{\mathcal{X}_j\times\mathcal{X}_{-j}}\left\|\nabla_j V(x_j, x_{-j})\right\|^2\mathrm{d}\rho_j\mathrm{d}\rho_{-j} = \left\|\nabla_{\mathrm{W}_2}\mathcal{V}(\rho)\right\|_{L^2(\rho)}^2.\end{aligned} \tag{18}$$

Thus in our proof, different from the Euclidean case, we cannot directly calculate the norm of the blockwise Wasserstein gradient of the updated coordinate and take expectation with respect to the index of the updated coordinate due to the same non-tensorization issue described in (18). We instead prove a high probability bound of the norm of blockwise Wasserstein gradient to control the functional value of $\mathcal{F}$ after each iteration. When this bound does not hold, we directly use the non-increasing property of the functional value to bound $\mathcal{F}$. By tuning the number M of updates in each iteration, we can derive the convergence rate of WPCG-R in Theorem 15. It is still an open question whether the iteration complexity in the theorem can be improved by applying other strategies. □

4.2 Convergence Rates of WPCG without λ-QG Assumption

When the λ-QG condition is not met for a strictly positive λ, we obtain the following convergence result for all three update schemes. This result parallels that for the first-order optimization algorithm in Euclidean spaces under convexity but not strict convexity.

Theorem 18 (**Polynomial convergence rate of WPCG without λ-QG**). Let $D_{\mathcal{X}} < \infty$ be the diameter of $\mathcal{X}$. Assume the step size $0 < \tau < \frac{m-1/2}{L(m-1)^{3/2}}$ if the coordinate is updated with the parallel scheme, and Assumptions A, B, and C hold. Then, for each update scheme, there exists a constant $C > 0$ depending on τ, m, L, and the update scheme, such that

$$\mathbb{E}\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{CD_{\mathcal{X}}^2 + \mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)}{k}.$$

In the parallel and the sequential update schemes, $\mathbb{E}\mathcal{F}(\rho^k) = \mathcal{F}(\rho^k)$ since there is no randomness.

Remark 19 (**Comparison with existing results in the Euclidean case**). When the QG condition is not met for some $\lambda > 0$, the functional value in a typical first-order optimization algorithm decreases with a rate no faster than $O(k^{-1})$ in the worst case. This has been demonstrated through various studies on the Euclidean space especially for the sequential update scheme. Theorem 3 in (Wright, 2015) shows that the function value decreases with a rate $O(k^{-1})$ when applying the coordinate GD with the sequential update scheme. Theorem 11.18 in (Beck, 2017) proves the same $O(k^{-1})$ convergence rate for the coordinate proximal GD with the sequential update scheme. Here, we derive the same convergence rate for WPCG with all three update schemes. □

The assumption of compactness of the parameter space is frequently utilized when applying the Wasserstein proximal gradient scheme to optimization problems, as seen in works by Santambrogio (2015); Yao and Yang (2022). This assumption plays a crucial role in our proof, allowing us to derive a uniform bound for $\mathrm{W}_2^2(\rho^k, \rho^*)$ and the first-order optimality condition of Wasserstein proximal gradient scheme (Lemma 9). The compactness is specifically used to determine the first variation of $\mathrm{W}_2^2(\mu, \nu)$ with respect to μ. More details on this can be found in Proposition 7.17 and Theorem 1.52 in (Santambrogio, 2015). It would be interesting to explore potential methods for relaxing this technical condition.

4.3 Convergence Rates of Inexact WPCG

In this subsection, we investigate the inexact WPCG algorithm, where non-zero numerical errors are allowed in each iteration. Our focus will be on the parallel update scheme (WPCG-P), though analogous convergence results can be derived similarly for the other two update schemes.

To precisely characterize the inexact WPCG-P algorithm, let $\widetilde{\rho}_j^{k+1}$ denote the inexact solution of the subproblem (13) for updating the j-th coordinate in the $(k+1)$-th iteration. We define a vector-valued function

$$\eta_j^{k+1} = T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} - \tau\left[\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^{k+1}) + \nabla \int_{\mathcal{X}} W_j(\cdot, y)\,\mathrm{d}\widetilde{\rho}_j^{k+1}(y) + \nabla \int_{\mathcal{X}} W_j(y, \cdot)\,\mathrm{d}\widetilde{\rho}_j^{k+1}(y)\right]$$

as the first variation of the objective functional in the subproblem (13) evaluated at $\widetilde{\rho}_j^{k+1}$. If $\widetilde{\rho}_j^{k+1}$ is the exact solution of the subproblem (13), the first-order optimality condition (Lemma 9) implies $\eta_j^{k+1} = 0$. Therefore, we can employ $\|\eta_j^{k+1}\|_{L^2(\mathcal{X}_j;\widetilde{\rho}_j^{k+1})}$ to characterize the numerical error incurred while solving the subproblem. Let $\{\varepsilon_j^k : j \in [m], k \in \mathbb{Z}_+\}$ be a sequence of error tolerance levels such that

$$\int_{\mathcal{X}_j} \|\eta_j^{k+1}\|^2\,\mathrm{d}\widetilde{\rho}_j^{k+1} \leqslant (\varepsilon_j^{k+1})^2$$

for every $j \in [m]$ and $k \in \mathbb{N}$. This rigorous formulation allows us to quantify the impact of numerical errors stemming from the inexact subproblem solutions on the overall convergence behavior of the WPCG-P algorithm.

Theorem 20 (**Convergence rate of inexact WPCG-P under λ-QG**). Assume Assumptions A, B, C, and D hold for some $\lambda > 0$, and the step size satisfies $0 < \tau < \frac{1}{2L\sqrt{m-1}}$. Let $\widetilde{\rho}^k$ be the inexact solution of WPCG-P at the k-th iteration with error tolerance levels $\varepsilon^k = (\varepsilon_1^k, \ldots, \varepsilon_m^k)$.

(1) If $\|\varepsilon^k\| \leqslant \varepsilon\kappa^k$ for some $\kappa \in (0,1)$ and $\varepsilon > 0$, then there are positive constants $C_5 = C_5(\lambda, \tau, m, L) < 1$ and $C_6 = C_6(\lambda, \tau, m, L, \kappa)$ such that

$$\mathrm{W}_2^2(\widetilde{\rho}^k, \rho^*) \leqslant \frac{2}{\lambda}\big[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)\big] \leqslant C_5^k\big[\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big] + C_6\varepsilon^2 \max\{C_5, \kappa^2\}^{k+1}.$$

(2) If $\|\varepsilon^k\| \leqslant \varepsilon k^{-\alpha}$ for some $\varepsilon, \alpha \geqslant 0$, then there exists a constant $C_7 = C_7(\lambda, \tau, m, L, \alpha) > 0$ such that

$$\mathrm{W}_2^2(\widetilde{\rho}^k, \rho^*) \leqslant \frac{2}{\lambda}\big[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)\big] \leqslant C_5^k\big[\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big] + \frac{C_7\varepsilon^2}{k^{2\alpha}}.$$

Remark 21 (**Impact of numerical error**). This result elucidates the impact of numerical errors, highlighting how the rate of numerical error reduction influences the overall convergence behavior. Specifically, the convergence rate of WPCG-P will be hindered if the numerical error decreases at a polynomial rate across iterations. Conversely, the algorithm maintains exponential convergence if the numerical error decays exponentially. Furthermore, if $\kappa^2 \leqslant C_5$, the same dependence of the iteration count on the condition number $\frac{L}{\lambda}$ and the number of blocks m can be derived by carefully tracking the constant C_5.

The following result demonstrates that the convergence rate remains polynomial in the presence of a numerical error that decreases polynomially, when the λ-QG condition is not met for a strictly positive λ. Moreover, the iteration count depends on the rate at which the numerical error diminishes, and as anticipated, will be no less than the iteration count of the exact WPCG-P algorithm.

Theorem 22 (**Polynomial convergence rate of inexact WPCG-P without λ-QG**). Let $D_{\mathcal{X}} < \infty$ be the diameter of $\mathcal{X}$. Assume the step size satisfies $0 < \tau < \frac{1}{2L\sqrt{m-1}}$ and Assumptions A, B, and C hold. If $\|\varepsilon^k\| \leqslant \varepsilon k^{-\alpha}$ for some $\varepsilon, \alpha \geqslant 0$, there exists a constant $C_8 = C_8(\alpha, \tau, m, L, D_{\mathcal{X}}, \varepsilon) > 0$ such that

$$\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) \leqslant \frac{C_8}{k^{\min\{1,\alpha\}}}$$

holds for all $k \in \mathbb{Z}_+$.

5 Numerical Experiments

In this section, we demonstrate the application of our WPCG algorithms to approximate the posterior distribution via the mean-field variational inference in Example 1 and to compute the stationary distribution of the multi-species systems with cross-interaction in Example 2.

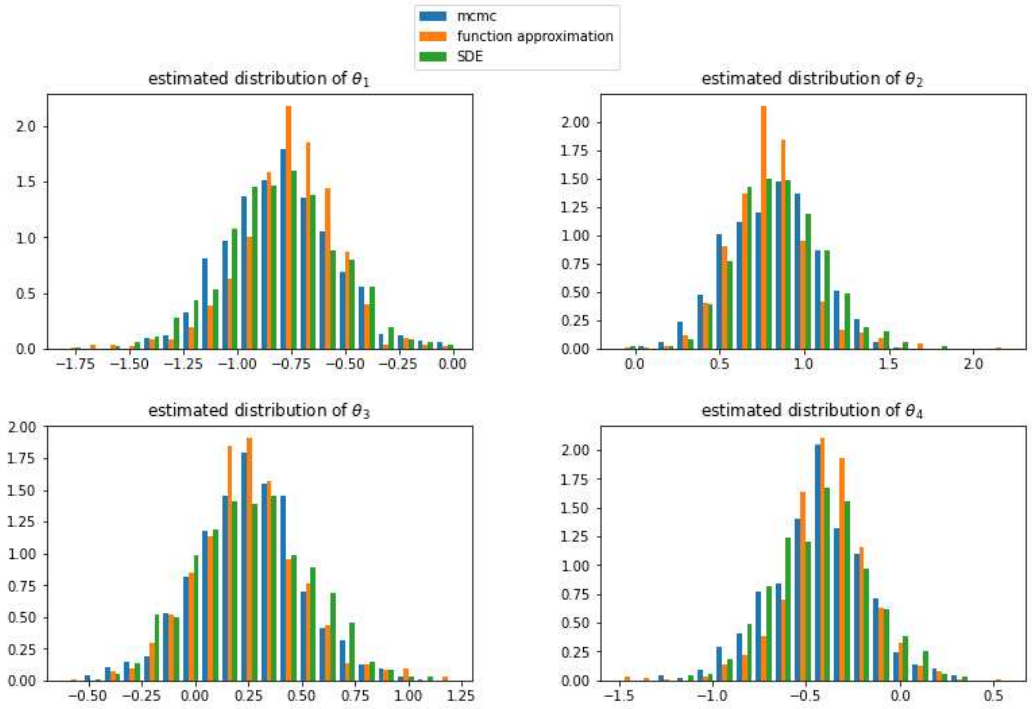


Figure 2: Histograms of particles used to approximate posterior distributions by MCMC and MFVI through WPCG-P with the FA approach and the SDE approach. Similar marginal distributions are derived by both WPCG-P (with the SDE approach or the FA approach) and MCMC.

5.1 Mean-field Variational Inference

Recall that $\Theta = \bigotimes_{j=1}^m \Theta_j$ is the parameter space, $p(x \,|\, \theta)$ is the likelihood function, π is the prior density function, and $X_1, \ldots, X_n$ are n observations. In MFVI, we approximate the posterior distribution by the solution of the following optimization problem,

$$\widehat{\pi}_n = \operatorname*{argmin}_{\rho = \bigotimes_{j=1}^m \rho_j} \mathrm{KL}(\rho \,\|\, \Pi_n), \quad \text{s.t.} \quad \rho_j \in \mathscr{P}(\Theta_j) \quad \forall\, j \in [m],$$

where Π_n is the posterior distribution with the density function given by (3). As a corollary of Theorem 9, we have the following convergence result about computing $\widehat{\pi}_n$ via WPCG-P.

Corollary 23. Let ρ^k be the k-th iterate by applying WPCG-P (Algorithm 1) to MFVI with $V = V_n = -\sum_{i=1}^n \log p(X_i \,|\, \theta) - \log \pi(\theta)$, $h_j(x) = x \log x$, and $W_j = 0$. If V_n is λ_n-strongly convex and L_n-smooth, and $0 < \tau < \frac{2m-1}{2L_n(m-1)^{3/2}}$, we have

$$\mathrm{W}_2^2(\rho^k, \widehat{\pi}_n) \leqslant (1 + C_n \lambda_n)^{-k} \big[\mathrm{KL}(\rho^0 \,\|\, \Pi_n) - \mathrm{KL}(\widehat{\pi}_n \,\|\, \Pi_n) \big],$$

where the constant

$$C_n = \frac{\frac{1}{2\tau} + (m-1)\big(\frac{1}{\tau} - L_n\sqrt{m-1}\big)}{2m\big[2L_n^2(m-1) + \frac{2}{\tau^2}\big]} > 0.$$

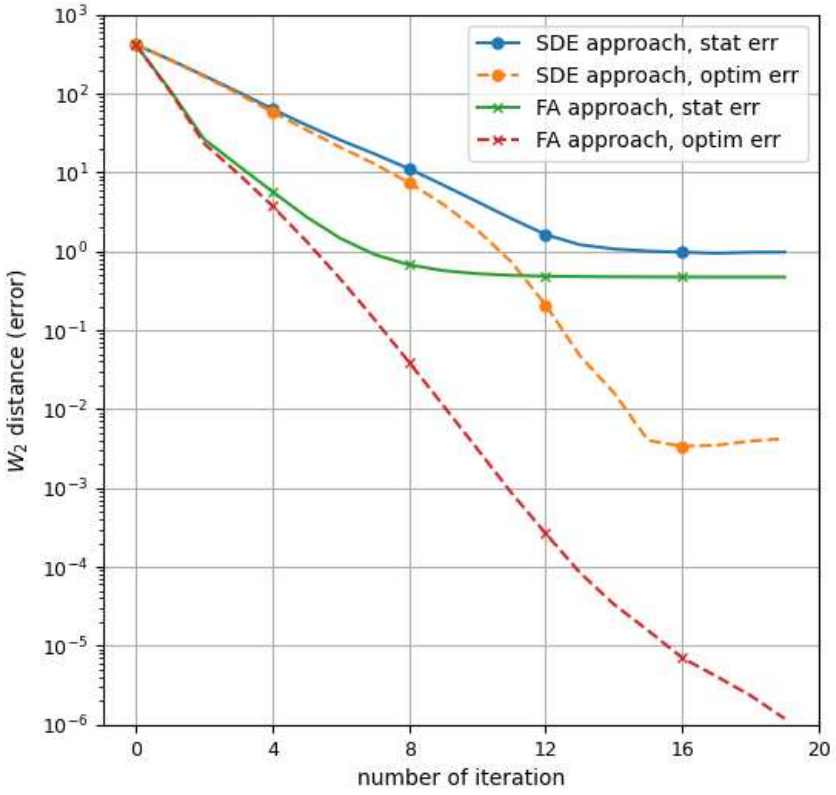


Figure 3: Numerical errors of the SDE approach and the FA approach. The numerical error $\mathrm{W}_2^2(\rho^k, \delta_{\theta^*})$ is first dominated by the optimization error $\mathrm{W}_2^2(\rho^k, \rho^*)$ which decays exponentially fast, and later dominated by the statistical error $\mathrm{W}_2^2(\rho^*, \delta_{\theta^*})$. The statistical error derived by the SDE approach is a bit larger than the one derived by the FA approach since we discretize the Langevin dynamics to approximate the Wasserstein proximal gradient scheme. The optimization error derived by the FA approach decreases exponentially fast as predicted by our theoretical results. In contrast, in the SDE approach, the optimization error will be dominated by the approximation error after several iterations.

Note that in Corollary 23, both λ_n and L_n depend on the samples $X_1, \ldots, X_n$. When the log-likelihood function $\log p(x \,|\, \theta)$ satisfies certain mild conditions (e.g., Assumptions 2 and 3 in Mei et al., 2016), the sample convexity parameter λ_n and the smoothness parameter L_n will concentration in the small neighborhoods of the population convexity parameter $\lambda = \mathbb{E}_{\theta^*} \lambda_n$ and the smoothness parameter $L = \mathbb{E}_{\theta^*} L_n$ respectively with high probability.

Remark 24. When applying WPCG-P to MFVI, our method coincides with representations in (Ghosh et al., 2022). However, Ghosh et al. (2022) only studied the convergence of the piecewise constant interpolation $\{\rho_t = \rho^{\lfloor t/\tau \rfloor} : t \geqslant 0\}$ towards the solution of its corresponding WGF (Equation (15) in Ghosh et al., 2022) as the step size $\tau \to 0$ when V_n is strongly convex. Yao and Yang (2022) proposed the MF-WGF algorithm for solving

MFVI with latent variables numerically. They showed exponential convergence of iterates towards the mean-field approximation $\widehat{\pi}_n$ of the true posterior in W_2 sense. Their proof heavily depended on the strong convexity of the negative log-likelihood function, while our analysis only uses the strong convexity to show QG condition. Lambert et al. (2022) studied Gaussian variational inference (i.e., $m = 1$ and the variational family is restricted to all Gaussian distributions) and showed that the corresponding gradient flow (so-called Bures–Wasserstein gradient flow) converges to the Gaussian variational approximation exponentially fast in W_2 sense. Compared with Gaussian variational inference, our method is more flexible with the number of blocks and the variational family. □

To support our theoretical findings, we simulate a Bayesian Logistic Regression model with prior $\pi = \mathcal{N}(0, 4I_4)$ and data generating process

$$X_i \sim \mathcal{N}(0, I_4), \quad y_i \,|\, X_i, \theta \sim \text{Bernoulli}(p_i) \quad \text{where} \quad \log \frac{p_i}{1-p_i} = X_i^T \theta,$$

with the ground truth being $\theta^* = (-1, 1, 0.3, -0.3)$. $n = 100$ samples generated from the true model are then used to estimate the posterior by implementing MFVI through the WPCG-P algorithm. Two different numerical methods, the FA approach and the SDE approach, are implemented. In both numerical methods, $B = 1000$ particles are used to approximate each marginal distribution. In the FA approach, a neural network with three fully connected hidden layers is used to approximate the optimal transport map. Each hidden layer consists of 1000 neurons, and the activation function is ReLu defined as $\mathrm{ReLu}(x) = \max\{x, 0\}$.

Figure 2 compares the WPCG-P algorithm with MCMC. The results show that similar approximations of each marginal distribution of the posterior distribution are derived by both methods. Figure 3 presents the numerical errors by applying the FA approach and the SDE approach in WPCG-P. The numerical error $\mathrm{W}_2^2(\rho^k, \delta_{\theta^*})$ is first dominated by the optimization error $\mathrm{W}_2^2(\rho^k, \rho^*)$ (corresponding to the decreasing parts of the solid lines) and then dominated by the statistical error $\mathrm{W}_2^2(\rho^*, \delta_{\theta^*})$ (corresponding to the horizontal parts of the solid lines) after several iterations. In the SDE approach, the optimization error will finally be dominated by the approximation error, which appears since we use discretized SDE to approximate the Wasserstein proximal gradient scheme, and stop decreasing as shown by the orange dashed line. As comparison, the optimization error in the FA approach will decrease exponentially fast as predicted by our theoretical findings, which is shown by the red dashed line.

5.2 Equilibrium in Multi-species Systems

Here we apply WPCG to find the stationary distribution of the following non-local multi-species cross-interaction model with diffusion, whose aggregation equation is given in (5). Our numerical result below shows that the algorithm converges exponentially fast as predicted by our theory when the corresponding objective functional is convex and satisfies the QG condition. We also conduct a numerical study to compare the FA approach and the SDE approach when $\mathcal{H}_j(\rho_j) = \int \rho_j \log \rho_j$ for all $j \in [m]$. The numerical study shows that the FA approach is not affected by the existence of super-quadratic terms in the potential energy function.

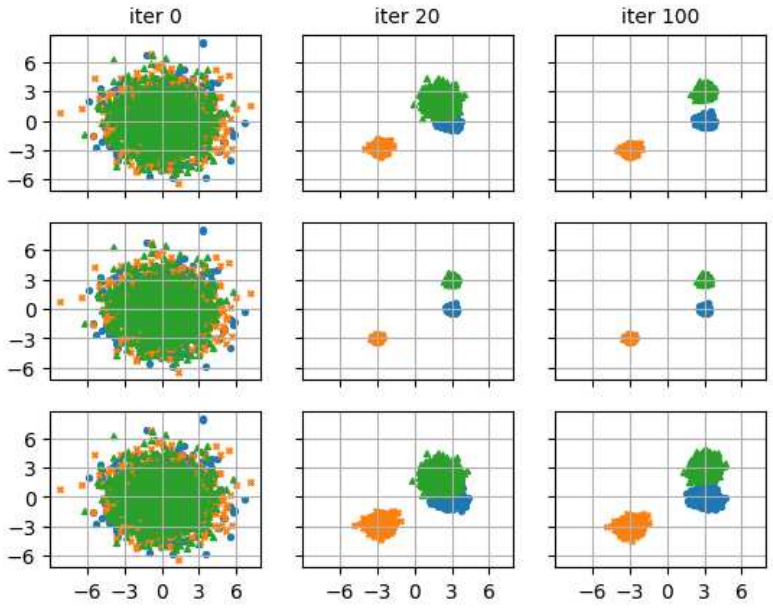


Figure 4: Scatter plots of all three species for different α and β. We use $B = 1000$ particles to approximate each species. The first row corresponds to $\alpha = 1$ and $\beta = 1$; the second row corresponds to $\alpha = 5$ and $\beta = 1$; the third row corresponds to $\alpha = 1$ and $\beta = 10$. Species concentrate more around the centers θ_1, θ_2, and θ_3 with larger α and smaller β.

Let us start from the following result which connects the PDE (5) and the optimization problem (1).

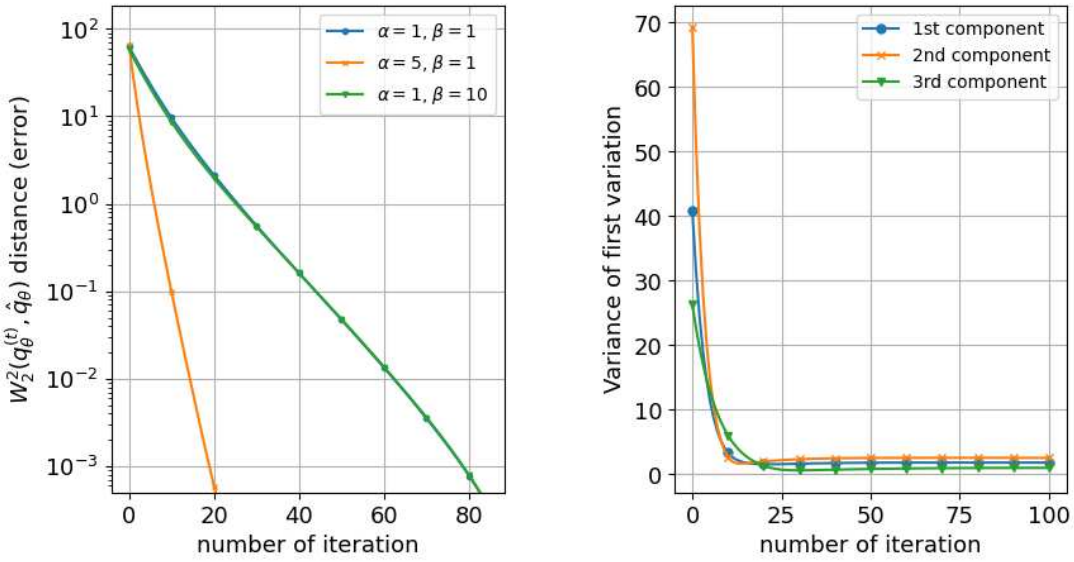
Proposition 25. Consider the optimization problem (1) with

$$V(x_1, \ldots, x_m) = \sum_{i=1}^{m} V_i(x_i) - \sum_{1 \leqslant i < j \leqslant m} K_{ij}(x_i - x_j)$$
$$W_j(x_j, x_j') = K_{jj}(x_j - x_j')$$

If Assumptions A, B, C, and D hold, and $K_{ij} = K_{ji}$ are even functions, then the solution $\rho^* = (\rho_1^*, \ldots, \rho_m^*)$ of problem (1) is the stationary distribution of the system (5).

Remark 26. (i) In this lemma, we assume $K_{ij} = K_{ji}$ for all $i, j \in [m]$ so that the evolution of $\{\rho_j(\cdot, t) : j \in [m], t \geqslant 0\}$ follows the WGF of some functional $\mathcal{F}$. When $m = 2$, this condition can be relaxed to $K_{12} = \alpha K_{21}$ for some constant $\alpha > 0$ by rescaling the WGF of ρ_1 and ρ_2 (Di Francesco and Fagioli, 2013). (ii) The choice of $h_j(x) = x \log x$ corresponds to the standard diffusion without medium of species j, and (5) turns to be the PDE of multi-species stochastic interacting particle systems (Daus et al., 2022). As a comparison, $h_j(x) = x^{m_j}$[2] corresponds to a porous medium type diffusion (Otto, 2001). When $m = 1$,

2. Here the exponent $m_j > 1$ depends on physical properties of the diffusing material.



(a) Optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$.

(b) $\mathrm{Var}_{\rho_j^k}\big(\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)(X_j^k)\big)$ for $1 \leqslant j \leqslant 3$.

Figure 5: (a) Comparison of optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$ for different α and β. All these lines are straight, indicating that the optimization error decays exponentially fast. Contraction rate does not change by only changing β (blue line v.s. green line), while it gets smaller (i.e., faster convergence) when α gets larger (orange line v.s. blue line). (b) Variance of first variation $\mathrm{Var}_{\rho_j^k}\big(\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)(X_j^k)\big)$ versus number of iterations. That the variance converges to 0 indicates that the first variation $\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)$ converges to a constant, which implies that $(\rho_1^k, \rho_2^k, \rho_3^k)$ converges to the minimum of $\mathcal{F}$.

(5) is the aggregation equation of interacting particle systems (Cabrales et al., 2020; Carrillo et al., 2019; Li and Rodrigo, 2010); when $m = 2$ and $V_j = h_j = 0$, (5) turns into the PDE of non-local interaction systems with two species (Di Francesco and Fagioli, 2013; Evers et al., 2017). To our knowledge, Equation (5) has not been studied for general $m \in \mathbb{N}^*$ and functions h_j, $j \in [m]$. We believe that our theory is also useful to study the existence and the uniqueness of this kind of PDEs from the perspective of WGF. □

In the numerical study, we consider the following test example (Daus et al., 2022; Jin et al., 2020) with $m = 3$ species in $\mathbb{R}^2$. We specify the functions

$$K_{ij}(x) = \frac{Q_i Q_j}{2} \arctan \|x\|^2, \quad h_j(\rho_j) = \beta \rho_j^2, \quad \text{and} \quad V_j(x_j) = \frac{\alpha r_j}{2} \|x_j - \theta_j\|^2, \quad 1 \leqslant i, j \leqslant 3,$$

with model parameters $(Q_1, Q_2, Q_3) = (1, -1, 0.5)$, $(r_1, r_2, r_3) = (6, 7, 3)$, and $\theta_1 = (3, 0)$, $\theta_2 = (-3, -3)$, and $\theta_3 = (3, 3)$. It is easy to check that (see Appendix D.2 for more details)

$$V(x_1, x_2, x_3) = \sum_{j=1}^{n} \frac{\alpha r_j}{2} \|x_j - \theta_j\|^2 - \sum_{1 \leqslant i < j \leqslant 3} \frac{Q_i Q_j}{2} \arctan \|x_i - x_j\|^2$$

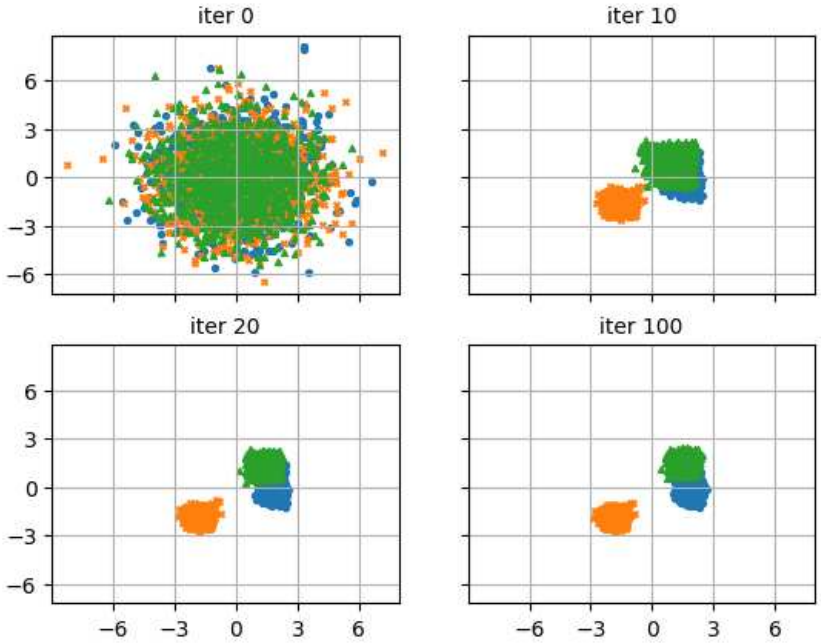


Figure 6: Scatter plots of all three species when the potential V contains a super quadratic term $\|x\|^4/4$. As expected, all species converge to their equilibrium; they get closer to the origin compared with the case without the super quadratic term.

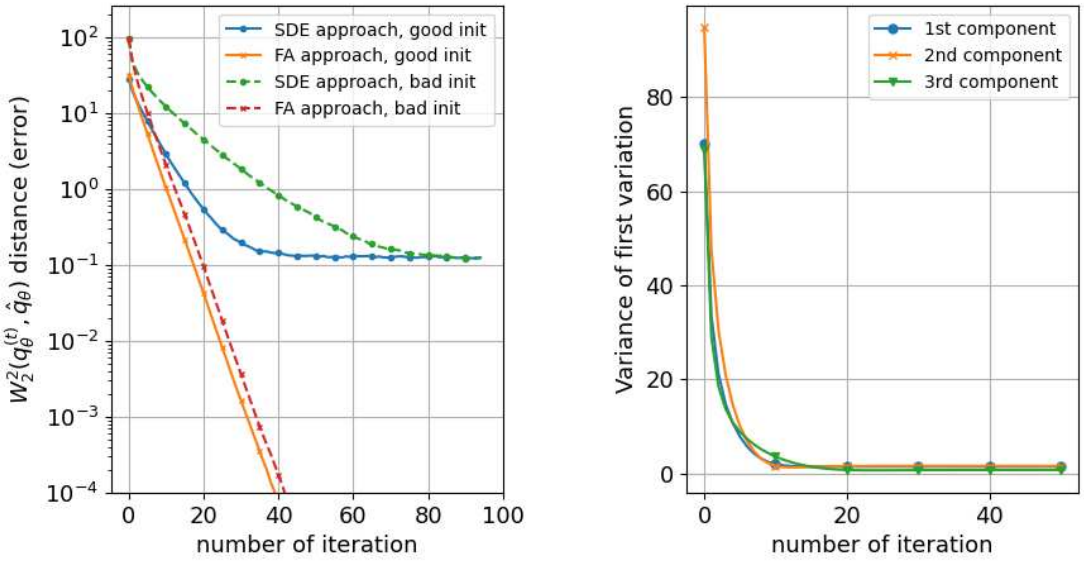
and

$$W_j(x_j, x_j') = -\frac{Q_j^2}{4} \arctan \|x_j - x_j'\|^2, \quad j = 1, 2, 3$$

satisfy Assumptions A, B, C, and D in Section 3 when $\alpha \geqslant 1$ and $\beta \geqslant 0$. We will apply WPCG with the FA approach by solving the optimization problem (16), since the corresponding $\mathcal{H}_j$ is not the negative self-entropy functional.

Figures 4, 5, 6, and 7 present the numerical results of this example. $B = 1000$ particles are generated to approximate each species, and a neural network with 2 hidden layers is used to solve (16) in the FA approach. Each hidden layer consists of 800 neurons, and the activation function is ReLu. In Figures 6 and 7, we add a super-quadratic term $\|x\|^4/4$ to the potential function V and change the entropy functional to the negative self-entropy to compare the SDE approach and the FA approach.

Figure 4 shows the scatter plots of all species with different α and β. When α gets larger, all species concentrate more around the corresponding centers (θ_1, θ_2, and θ_3) due to a larger external force. In contrast, larger β makes all species less concentrate since the entropy term penalizes the concentration of species. Figure 5 shows the optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$ for different α and β and the variance of the first variation when $\alpha = \beta = 1$. In Figure 5a, the optimization error decays faster with larger α (compare the blue line with the orange line) since the convexity gets stronger, while changing β will not change the convergence speed (compare the blue line with the green line) since the entropy term does



(a) Optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$.

(b) $\mathrm{Var}_{\rho_j^k}\left(\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)(X_j^k)\right)$ for $1 \leqslant j \leqslant 3$.

Figure 7: (a) Comparison of optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$ derived by the SDE approach and the FA approach when $h_j(x) = x \log x$. The optimization error derived by applying the FA approach (the orange line and the red line) decays exponentially fast despite of the existence of the super-quadratic term. In contrast, the optimization error derived by applying the SDE approach (the green line and the blue line) is dominated due to the approximation error. Moreover, due to the super-quadratic term, it takes more iterations for the SDE approach to converge when the initialization gets farther from the origin. (b) Variance of the first variation $\mathrm{Var}_{\rho_j^k}\left(\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)(X_j^k)\right)$ versus number of iterations. That the variance converges to 0 indicates that the first variation $\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)$ converges to a constant, which implies that $(\rho_1^k, \rho_2^k, \rho_3^k)$ converges to the minimum of $\mathcal{F}$.

not provide any convexity to the whole functional. Figure 5b plots the variance of the first variation $\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)(X_j^k)$ where $X_j^k \sim \rho_j^k$. This variance converges to 0 if and only if the first variation $\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)$ converges to a constant $[\rho_j]$-a.e., which indicates that $(\rho_1^k, \rho_2^k, \rho_3^k)$ converges to the minimum of $\mathcal{F}$.

Figure 6 shows the scatter plot of all species when there exists an extra super-quadratic term $\|x\|^4/4$ in the potential function V. Due to this super-quadratic term, all species get closer to the origin when the system is at its equilibrium. Figure 7 shows the optimization error $\mathrm{W}_2^2(\rho_\theta^k, \rho_\theta^*)$ and the variance of the first variation. Figure 7a compares the optimization error derived by WPCG with the FA approach and the SDE approach. Compare with the FA approach, the optimization error derived by applying the SDE approach (the blue line and the green line) is dominated by the approximation error after several iterations. Due to the existence of the super-quadratic term, a smaller step size has to be chosen to make the

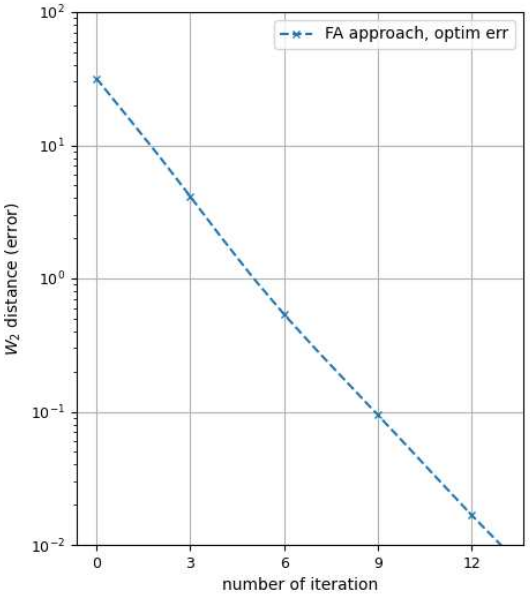


Figure 8: Optimization error $\mathrm{W}_2^2(\rho^k, \rho^*)$.

SDE approach converge when the initialization gets farther from the origin, which increases the number of iterations for the SDE approach to converge. In contrast, the optimization error derived by the FA approach (the red line and the orange line) is not affected by the super-quadratic term as predicted by our theoretical result. Fig 7b plots the variance of the first variation $\frac{\delta \mathcal{F}}{\delta \rho_j}(\rho^k)$. The variance converges to zero as a sign of convergence of ρ^k to the minimum of $\mathcal{F}$.

5.3 Real Data Example

We apply the WPCG-P algorithm to analyze the Pima Indian diabetes data set (Smith et al., 1988). This data set comprises 8 medical predictors (`pregnancies`, `glucose`, `blood pressure`, `skin thickness`, `insulin`, `body mass index`, `diabetes pedigree function`, `age`) and a binary outcome variable (0 for no diabetes, 1 have diabetes) from 768 individuals.

We construct a Bayesian logistic regression model (with an intercept term) to predict the outcome based on all 8 medical predictors. A stratified sample of 68 data points is set aside as the test set, while the remaining 700 points constitute the training set. We employ the WPCG-P algorithm with the FA approach to compute the mean-field variational approximation of the posterior distribution over the intercept and predictor coefficients. Figure 8 depicts the optimization error $\mathrm{W}_2^2(\rho^k, \rho^*)$ across iterations, which decays exponentially fast as predicted by our theoretical findings.

Furthermore, We compare our algorithm with MCMC for sampling from the posterior distribution. Both WPCG-P and MCMC achieve a misclassification rate 19.1% and a cross entropy loss 0.496 on the test set. Figure 9 presents the box plot of the mean-field

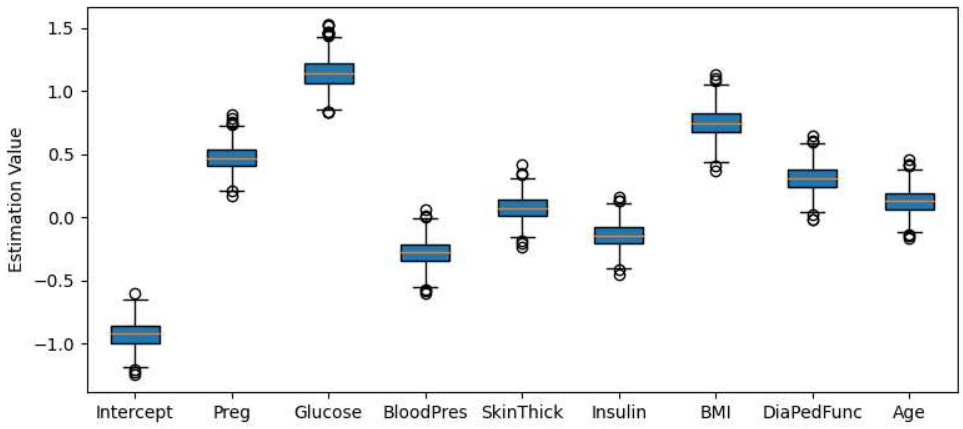


Figure 9: Box plot of the mean-field variational approximation of the posterior distribution of the intercept and all predictors.

variational approximations of the posterior distributions over the intercept and predictor coefficients. Consistent with domain knowledge, the analysis indicates that larger values of `pregnancies`,`Glucose`, `BMI`, and `diabetes pedigree function` are associated with an increasing possibility of having diabetes.

6 Conclusions and Discussions

In this paper, we introduced the WPCG algorithms for solving multivariate composite convex minimization problem on the Wasserstein space. We established its exponential convergence results with different update schemes under the quadratic growth condition. When this condition is not met, we also showed a slower polynomial convergence result for the WPCG algorithm. Additionally, we analyze the convergence behavior of the inexact WPCG algorithm with the parallel update scheme. We believe that similar results hold for the other two update schemes. We also conducted numerical studies about mean-field variation inference and multi-species systems to verify the predictions from our theory.
To conclude this paper, we list several open problems as future directions.

1. An interesting topic is to develop the PL-type condition for multivariate objective functionals and investigate its relationship with the QG condition. This will enable us to relax the assumptions of convexity and the QG condition, and analyze the convergence rate of the WPCG algorithm for non-convex objective functionals.

2. It is interesting to see whether it would be beneficial to augment the Euclidean space $\mathbb{R}^m$ to the space of all product measures $\bigotimes_{j=1}^m \rho_j$, which includes $\mathbb{R}^m$ as a special case by restricting each ρ_j to be a point mass measure, when minimizing a function V over $\mathbb{R}^m$. This augmentation may help avoid the algorithm quickly getting trapped into a local minimum in case of non-convexity, similar to SGD which avoids getting

trapped by injecting random noise; the augmented algorithm can be viewed as the infinite particle limit of an evolutionary algorithm for minimizing V.

3. It is still unknown if the convergence rates of WPCG are optimal with respect to the number of blocks m. The dependence on m will be important when m is very large. Also, The current convergence result of WPCG-R seems worse than the one in the coordinate GD with the random update scheme on the Euclidean space. It is interesting to see if this convergence rate is due to the analyzing strategies or the difference between the Wasserstein space and the Euclidean space.

4. Another potential topic is to consider the coordinate Wasserstein forward-backward gradient descent algorithm, which updates each coordinate by approximating the potential energy $\mathcal{V}$ in first order (forward step) followed by a proximal descent step (backward step). This algorithm corresponds to the proximal coordinate gradient descent algorithm in the Euclidean case. The case $m = 1$ on the Wasserstein space has been studied by Salim et al. (2020). We believe our analysis of WPCG can be extended to the coordinate Wasserstein forward-backward gradient descent algorithm.

5. In many problems in statistics (such as MFVI in Section 5.1), the potential function takes the form $V = \sum_{i=1}^{n} V_i$, where n can be treated as the sample size. When n is large, there will be huge computational cost to calculate ∇V in the proximal gradient descent step. One possible way is to approximate $\nabla V \approx \frac{n}{n_s} \sum_{l=1}^{n_s} \nabla V_{i_l}$ by subsampling i_l from $[n] := \{1, \ldots, n\}$ independently for $l \in [n_s]$ in each update, where n_s is the batch size. It is interesting to see how this stochastic WPCG algorithm converges and the impact of the batch size n_s.

Acknowledgments and Disclosure of Funding

Xiaohui Chen was partially supported by NSF CAREER grant DMS-2347760, NSF grant DMS-2413404, and a gift from the Simons Foundation. Yun Yang was partially supported by NSF grant DMS-2210717.

Appendices

Appendix A. Proof of Main Results

Since the whole proof is quite involved, we will first provide with the sketch, and detail the proof in next several subsections. For every update scheme when $\lambda > 0$, we proceed with the following two steps

Step 1. (Sufficient decrease.) In this step, we will prove

$$\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) \geqslant C\, \mathrm{W}_2^2(\rho^{k-1}, \rho^k)$$

holds with some constant $C > 0$ depending only on τ, λ, L, and the update scheme.

Step 2. (Bound of functional value.) In this step, we will prove

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant C'\, \mathrm{W}_2^2(\rho^k, \rho^{k-1})$$

holds with some constant $C' > 0$ depending only on τ, λ, L, and the update scheme.

With the above two steps, we have

$$\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) \geqslant C\, \mathrm{W}_2^2(\rho^{k-1}, \rho^k) \geqslant \frac{C}{C'}\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big)$$

This implies

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \Big(1 + \frac{C}{C'}\Big)^{-1}\big(\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big) \leqslant \cdots \leqslant \Big(1 + \frac{C}{C'}\Big)^{-k}\big(\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big).$$

Since $\mathcal{F}$ satisfies (λ-QG) condition, we have

$$\frac{\lambda}{2}\, \mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \Big(1 + \frac{C}{C'}\Big)^{-k}\big(\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big).$$

This implies the desired result. Next, we will detail the proof for different update schemes.

A.1 Parallel Update Scheme: Proof of Theorem 9

As introduced in the sketch, we will use the following two results to bound the decrease of functional values and the difference between the functional value and the minimum of $\mathcal{F}$, whose proof will be postponed to Appendix B.

Lemma 1 (sufficient decrease). Under Assumptions A, B, and C, when the step size satisfies $0 < \tau < \frac{2m-1}{2L(m-1)^{3/2}}$, we have

$$\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) \geqslant C_4(m, L, \tau)\, \mathrm{W}_2^2(\rho^{k-1}, \rho^k) > 0,$$

where the constant

$$C_4(m, L, \tau) = \frac{1}{m}\bigg[\frac{1}{2\tau} + (m-1)\Big(\frac{1}{\tau} - L\sqrt{m-1}\Big)\bigg].$$

Lemma 2 (bound of functional value). Under Assumptions A, B, C, and D, in both the parallel update scheme and the sequential update scheme, we have

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{2}{\lambda}\Big(2L^2(m-1) + \frac{2}{\tau^2}\Big)\,\mathrm{W}_2^2(\rho^k, \rho^{k-1}).$$

With the above two lemmas,

$$\begin{aligned}\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) &\overset{\text{Lem 1}}{\geqslant} \frac{1}{m}\bigg[\frac{1}{2\tau} + (m-1)\Big(\frac{1}{\tau} - L\sqrt{m-1}\Big)\bigg]\,\mathrm{W}_2^2(\rho^{k-1}, \rho^k)\\ &\overset{\text{Lem 2}}{\geqslant} \frac{\frac{1}{m}\big[\frac{1}{2\tau} + (m-1)(\frac{1}{\tau} - L\sqrt{m-1})\big]}{\frac{2}{\lambda}\big(2L^2(m-1) + \frac{2}{\tau^2}\big)}\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big)\\ &= \lambda C_1(m, L, \tau)\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big),\end{aligned}$$

where

$$C_1(m, L, \tau) = \frac{\frac{1}{2\tau} + (m-1)\big(\frac{1}{\tau} - L\sqrt{m-1}\big)}{2m[2L^2(m-1) + \frac{2}{\tau^2}]}$$

is defined in Theorem 9. Therefore, we have

$$\frac{\lambda}{2}\,\mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \big[1 + \lambda C_1(m, L, \tau)\big]^{-t}\big(\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big),$$

where the first inequality is due to Assumption D (λ-QG condition). This ends the proof.

A.2 Sequential Update Scheme: Proof of Theorem 12

In the sequential update scheme, we have the following lemma to bound the decrease of functional values, whose proof will be postponed to Appendix B.

Lemma 3 (sufficient decrease). For two consecutive updates ρ^{k-1} and ρ^k, we have

$$\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) \geqslant \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho^{k-1}, \rho^k).$$

With the above lemma, we have

$$\begin{aligned}\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) &\overset{\text{Lem 3}}{\geqslant} \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho^{k-1}, \rho^k)\\ &\overset{\text{Lem 2}}{\geqslant} \frac{\lambda}{4\tau}\Big(2L^2(m-1) + \frac{2}{\tau^2}\Big)^{-1}\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big)\\ &= \lambda C_2(m, L, \tau)\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big).\end{aligned}$$

Therefore, we have

$$\frac{\lambda}{2}\,\mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \mathcal{F}(\rho^k, \rho^*) \leqslant \big[1 + \lambda C_2(m, L, \tau)\big]^{-k}\big(\mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)\big),$$

where the first inequality is due to Assumption D (λ-QG condition). This ends the proof.

A.3 Random Update Scheme: Proof of Theorem 15

For any fixed $k \in \mathbb{N}$, let $\mathcal{A}_k^T(M)$ be the event such that all coordinates are updated at least once but at most T times in the k-th iteration. Take

$$M_\delta = \left\lceil m \log \frac{m}{\delta} \right\rceil \quad \text{and} \quad T_\delta = \left\lceil e \log \frac{m}{\delta} \right\rceil.$$

The following lemma controls the probability of $\mathcal{A}_k^{T_\delta}(M_\delta)^c$, whose proof will be postponed to Appendix B.

Lemma 4. For any $\delta > 0$ such that $0 < \delta < 1$, define

$$M = \left\lceil m \log \frac{m}{\delta} \right\rceil \quad \text{and} \quad T = \left\lceil e \log \frac{m}{\delta} \right\rceil.$$

$j_0, \ldots, j_{M-1}$ are i.i.d. random variables sampled from Uniform$\{1, \ldots, m\}$. Then, there is probability at least $1 - 2\delta$, such that every integer in $\{1, 2, \ldots, m\}$ appears at least once but at most T times in $\{j_l\}_{l=0}^{M-1}$.

From the above lemma, we know $\mathbb{P}(\mathcal{A}_t^{T_\delta}(M_\delta)^c) \leqslant 2\delta$. Note that

$$\begin{aligned}\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \rho^{k-1}\big] = {} & \mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \mathcal{A}_k^{T_\delta}(M_\delta), \rho^{k-1}\big] \cdot \mathbb{P}(\mathcal{A}_k^{T_\delta}(M_\delta)) \\ & + \mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \mathcal{A}_k^{T_\delta}(M_\delta)^c, \rho^{k-1}\big] \cdot \mathbb{P}(\mathcal{A}_k^{T_\delta}(M_\delta)^c).\end{aligned}$$

The second term can be bounded by $2\delta\big(\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big)$ since $\mathcal{F}(\rho^k)$ is non-increasing with respect to k. Next, we will bound the first term with the following two lemmas, whose proof will be postponed to Appendix B.

Lemma 5 (sufficient decrease). For any $k, M \in \mathbb{Z}_+$, we have

$$\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) \geqslant \frac{1}{2\tau} \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}).$$

Lemma 6 (bound of functional value). For any $k \in \mathbb{Z}_+$, let $M \in \mathbb{Z}_+$ be an integer such that all coordinates are updated at least once but at most T times through M updates in the k-th iteration. Under Assumptions A, B, C, and D, we have

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{4}{\lambda}\Big(L^2 m T + \frac{1}{\tau^2}\Big) \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}).$$

Under $\mathcal{A}_k^{T_\delta}(M_\delta)$, we have

$$\begin{aligned}\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) & \overset{\text{Lem 5}}{\geqslant} \frac{1}{2\tau} \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}) \\ & \overset{\text{Lem 6}}{\geqslant} \frac{\lambda}{8\tau}\Big(L^2 m T_\delta + \frac{1}{\tau^2}\Big)^{-1}\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big).\end{aligned}$$

This implies

$$\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \mathcal{A}_k^{T_\delta}(M_\delta), \rho^{k-1}\big] \leqslant \Big(1 + \frac{\lambda\tau}{8(1 + \tau^2 L^2 m T_\delta)}\Big)^{-1}\Big(\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\Big).$$

Combining all the pieces above yields

$$\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \rho^{k-1}\big] \leqslant \left[\Big(1 + \frac{\lambda\tau}{8(1+\tau^2L^2mT_\delta)}\Big)^{-1} + 2\delta\right]\big(\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big)$$

Taking the expectation with respect to ρ^{k-1}, we have

$$\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big] \leqslant \left[\Big(1 + \frac{\lambda\tau}{8(1+\tau^2L^2mT_\delta)}\Big)^{-1} + 2\delta\right]\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big].$$

Now, let us choose a proper δ to derive the convergence rate. Taking $\delta = 1/(mL^2)$, we have $T_\delta = 2e\log(mL)$ and $M_\delta = \big\lceil 2m\log(mL)\big\rceil$. Therefore, we have

$$\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big] \leqslant C_3(m, L, \tau)\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big]$$

where

$$C_3(m, L, \tau) = \left(1 + \frac{\lambda\tau}{8[1 + 2e\tau^2L^2m\log(mL)]}\right)^{-1} + \frac{2}{mL^2},$$

Note that this is a decreasing algorithm, meaning that $\mathcal{F}(\rho^k) \leqslant \mathcal{F}(\rho^{k-1})$. Therefore,

$$\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big] \leqslant \min\big\{C_3(m, L, \tau), 1\big\}\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big].$$

Taking $\tau^{-1} = \sqrt{2emL^2\log(mL)}$ yields

$$C_3 = \left(1 + \frac{\lambda}{2\sqrt{2eL^2m\log(mL)}}\right)^{-1} + \frac{2}{mL^2}.$$

A.4 Case $\lambda = 0$: Proof of Theorem 18

Parallel and sequential update schemes. First, let us show there is a constant C depending on m, L, τ and the update scheme, such that

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant C\sqrt{\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k)}\,\mathrm{W}_2(\rho^k, \rho^*).$$

In fact, we have

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \overset{\text{(i)}}{\leqslant} C\,\mathrm{W}_2(\rho^k, \rho^{k-1})\,\mathrm{W}_2(\rho^k, \rho^*) \overset{\text{(ii)}}{\leqslant} C\sqrt{\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k)}\cdot\mathrm{W}_2(\rho^k, \rho^*),$$

where C is a constant varying from line to line. Here, (i) is by the inequality (24) in both the parallel and the sequential update schemes; (ii) is by Lemma 1 in the parallel update scheme and Lemma 3 in the sequential update scheme.

The above inequality implies

$$\begin{aligned}\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*) &\geqslant \big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big) + \frac{1}{C\,\mathrm{W}_2^2(\rho^k, \rho^*)}\cdot\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big)^2\\ &\geqslant \big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big) + \frac{1}{CD_{\mathcal{X}}^2}\cdot\big(\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big)^2.\end{aligned}$$

In the last inequality, we use the fact that

$$\mathrm{W}_2^2(\rho^t, \rho^*) = \inf_{X\sim\rho^k, Y\sim\rho^*} \mathbb{E}\|X - Y\|^2 \leqslant D_{\mathcal{X}}^2.$$

Therefore, we have

$$\frac{1}{\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)} - \frac{1}{\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)} \geqslant \frac{1}{CD_{\mathcal{X}}^2 + \mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)} \geqslant \frac{1}{CD_{\mathcal{X}}^2 + \mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)},$$

where the last inequality is due the non-increasing property of $\mathcal{F}$ derived by Lemma 1 and Lemma 3. This implies

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{CD_{\mathcal{X}}^2 + \mathcal{F}(\rho^0) - \mathcal{F}(\rho^*)}{k}.$$

Random update scheme. In the random update scheme, let $\mathcal{A}_k^T(M)$ be the event such that all coordinates are updated at least once but at most T times in the k-th iteration. We have $\mathbb{P}(\mathcal{A}_k^{T_\delta}(M_\delta)^c) \leqslant 2\delta$ by taking

$$M_\delta = \left\lceil m \log \frac{m}{\delta} \right\rceil \quad \text{and} \quad T_\delta = \left\lceil e \log \frac{m}{\delta} \right\rceil.$$

Similar to the proof of Theorem 15, we consider the decomposition

$$\begin{aligned}\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \rho^{k-1}\big] = {}& \mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \mathcal{A}_k^{T_\delta}(M_\delta), \rho^{k-1}\big] \cdot \mathbb{P}(\mathcal{A}_k^{T_\delta}(M_\delta)) \\ & + \mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \,\big|\, \mathcal{A}_k^{T_\delta}(M_\delta)^c, \rho^{k-1}\big] \cdot \mathbb{P}(\mathcal{A}_k^{T_\delta}(M_\delta)^c).\end{aligned}$$

By (25) and Lemma 5, the first term is bounded by

$$\sqrt{2\tau\Big(2L^2 m T_\delta + \frac{2}{\tau^2}\Big)} \cdot \mathcal{D}_{\mathcal{X}} \sqrt{\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k)}.$$

Since $\mathcal{F}(\rho^k)$ is non-increasing, the second term can be bounded by $2\delta\big(\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big)$. Thus, we have

$$\begin{aligned}&\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big] \\ &\leqslant \sqrt{2\tau\Big(2L^2 m T_\delta + \frac{2}{\tau^2}\Big)} \cdot \mathbb{E}\mathcal{D}_{\mathcal{X}} \sqrt{\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k)} + 2\delta\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big] \\ &\leqslant \sqrt{2\tau\Big(2L^2 m T_\delta + \frac{2}{\tau^2}\Big)} \cdot \mathcal{D}_{\mathcal{X}} \sqrt{\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k)\big]} + 2\delta\mathbb{E}\big[\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^*)\big],\end{aligned} \tag{19}$$

where the last inequality is due to Cauchy–Schwarz inequality.

Now, let us prove $\mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big] \lesssim \frac{1}{k}$. For simplicity, assume $4\delta \leqslant 1$ and let

$$x_k = \mathbb{E}\big[\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*)\big], \quad A_\delta = D_{\mathcal{X}} \sqrt{2\tau\Big(2L^2 m T_\delta + \frac{2}{\tau^2}\Big)}, \quad \text{and} \quad \tilde{A}_\delta = \frac{A_\delta}{1 - 4\delta}.$$

We will prove $x_k \leqslant (\tilde{A}_\delta^2 + x_0)/k$ for all $k \geqslant 1$ by induction. $k = 1$ is obvious since $x_1 \leqslant x_0$. For $k \geqslant 2$, note that (19) implies

$$x_k \leqslant A_\delta \sqrt{x_{k-1} - x_k} + 2\delta x_{k-1}. \tag{20}$$

Case 1: $x_{k-1} \leqslant 2x_k$. (20) implies $x_k \leqslant \tilde{A}_\delta \sqrt{x_{k-1} - x_k}$. Thus, we have

$$\frac{1}{x_k} \geqslant \frac{1}{x_k + \tilde{A}_\delta^2} + \frac{1}{x_{k-1}} \overset{\text{(i)}}{\geqslant} \frac{1}{x_0 + \tilde{A}_\delta^2} + \frac{k-1}{\tilde{A}_\delta^2 + x_0} = \frac{k}{x_0 + \tilde{A}_\delta^2}.$$

Here, (i) is by $x_k \leqslant x_0$ and the induction hypothesis.
Case 2: $x_{k-1} > 2x_k$. In this case, we have

$$x_k < \frac{x_{k-1}}{2} \overset{\text{(i)}}{\leqslant} \frac{\tilde{A}_\delta^2 + x_0}{2(k-1)} \overset{\text{(ii)}}{\leqslant} \frac{\tilde{A}_\delta^2 + x_0}{k}.$$

Here, (i) is due to the induction hypothesis, and (ii) is by $k \geqslant 2$. By induction, we have $x_k \leqslant (\tilde{A}_\delta^2 + x_0)/k$. We finish the proof.

A.5 Inexact WPCG-P with $\lambda > 0$: Proof of Theorem 20

The proof is similar to the exact WPCG, but we need to take the cumulative error into consideration. Some analogous lemmas as in the proof of Theorem 9 are required, the proofs of which are postponed to Appendix B.

Lemma 7 (sufficient decrease for inexact WPCG-P). Under Assumptions A, B, and C, when the step size satisfies $0 < \tau < \frac{1}{2L\sqrt{m-1}}$, we have

$$\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\widetilde{\rho}^{k+1}) \geqslant \Big(\frac{1}{2\tau} - L\sqrt{m-1}\Big)\, \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \frac{\|\varepsilon^{k+1}\|^2}{2\tau}.$$

Lemma 8 (bound of functional value for inexact WPCG-P). Under Assumptions A, B, C, and D, we have

$$\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) \leqslant \frac{6}{\lambda}\Big[\Big(L^2(m-1) + \frac{1}{\tau^2}\Big)\, \mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{\|\varepsilon^k\|^2}{\tau^2}\Big].$$

With the above two lemmas, we have

$$\begin{aligned}\frac{\lambda}{6}\big[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)\big] - \frac{\|\varepsilon^k\|^2}{\tau^2} &\overset{\text{(i)}}{\leqslant} \Big(L^2(m-1) + \frac{1}{\tau^2}\Big)\, \mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) \\ &\overset{\text{(ii)}}{\leqslant} \frac{L^2(m-1) + \frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}}\Big[\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\widetilde{\rho}^k) + \frac{\|\varepsilon^k\|^2}{2\tau}\Big].\end{aligned}$$

Here (i) is by Lemma 8, and (ii) is by Lemma 7. Reorganizing the inequality yields

$$\begin{aligned}\left[\frac{\lambda}{6} + \frac{L^2(m-1) + \frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}}\right]\big[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)\big] &\leqslant \frac{L^2(m-1) + \frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}}\big[\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\rho^*)\big] \\ &\quad + \left[\frac{L^2(m-1) + \frac{1}{\tau^2}}{1 - 2\tau L\sqrt{m-1}} + \frac{1}{\tau^2}\right]\|\varepsilon^k\|^2.\end{aligned}$$

Applying Lemma 14 with

$$A = \Big(1 + \frac{\lambda}{6} \cdot \frac{\frac{1}{2\tau} - L\sqrt{m-1}}{L^2(m-1) + \frac{1}{\tau^2}}\Big)^{-1}, \quad B = \frac{\frac{L^2(m-1)+\frac{1}{\tau^2}}{1-2\tau L\sqrt{m-1}} + \frac{1}{\tau^2}}{\frac{L^2(m-1)+\frac{1}{\tau^2}}{\frac{1}{2\tau}-L\sqrt{m-1}} + \frac{\lambda}{6}}, \quad \text{and} \quad \xi_k = \|\varepsilon^k\|^2$$

yields the convergence result.

A.6 Inexact WPCG-P with $\lambda > 0$: Proof of Theorem 22

Recall that we have Equation (26)

$$\begin{aligned}\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) &\leqslant \sqrt{\Big(3L^2(m-1) + \frac{3}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}} \cdot \mathrm{W}_2(\widetilde{\rho}^k, \rho^*)\\ &\leqslant \sqrt{3}\,\mathrm{W}_2(\widetilde{\rho}^k, \rho^*)\bigg[\frac{\|\varepsilon^k\|}{\tau} + \sqrt{L^2(m-1) + \frac{1}{\tau^2}}\,\mathrm{W}_2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1})\bigg]\\ &\leqslant \sqrt{3}D_{\mathcal{X}}\bigg[\frac{\|\varepsilon^k\|}{\tau} + \sqrt{\frac{L^2(m-1)+\frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}} \cdot \Big(\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\widetilde{\rho}^k) + \frac{\|\varepsilon^k\|^2}{2\tau}\Big)}\bigg].\end{aligned}$$

Here, we apply Lemma 7 and the fact that $\mathrm{W}_2(\widetilde{\rho}^k, \rho^*) \leqslant D_{\mathcal{X}}$. If $\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) \geqslant \sqrt{3}D_{\mathcal{X}}\frac{\|\varepsilon^k\|}{\tau}$, we have

$$\Big(\frac{\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)}{\sqrt{3}D_{\mathcal{X}}} - \frac{\|\varepsilon^k\|}{\tau}\Big)^2 \leqslant \frac{L^2(m-1)+\frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}}\Big(\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\widetilde{\rho}^k) + \frac{\|\varepsilon^k\|^2}{2\tau}\Big),$$

otherwise

$$\Big(\frac{\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)}{\sqrt{3}D_{\mathcal{X}}} - \frac{\|\varepsilon^k\|}{\tau}\Big)^2 \leqslant \frac{\|\varepsilon^k\|^2}{\tau^2}.$$

Therefore, we have

$$\Big(\frac{\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)}{\sqrt{3}D_{\mathcal{X}}} - \frac{\|\varepsilon^k\|}{\tau}\Big)^2 \leqslant \frac{L^2(m-1)+\frac{1}{\tau^2}}{\frac{1}{2\tau} - L\sqrt{m-1}}\Big(\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\widetilde{\rho}^k) + \frac{\|\varepsilon^k\|^2}{2\tau}\Big) + \frac{\|\varepsilon^k\|^2}{\tau^2},$$

i.e.

$$\begin{aligned}&\frac{\frac{1}{2\tau} - L\sqrt{m-1}}{L^2(m-1)+\frac{1}{\tau^2}} \cdot \frac{[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)]^2}{3D_{\mathcal{X}}^2} + \Big[1 - \frac{\frac{1}{2\tau} - L\sqrt{m-1}}{L^2(m-1)+\frac{1}{\tau^2}} \cdot \frac{2\|\varepsilon^k\|}{\sqrt{3}\tau D_{\mathcal{X}}}\Big] \cdot \big[\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*)\big]\\ &\qquad\leqslant \big[\mathcal{F}(\widetilde{\rho}^{k-1}) - \mathcal{F}(\rho^*)\big] + \frac{\|\varepsilon^k\|^2}{2\tau}.\end{aligned}$$

Applying Lemma 15 with

$$A = \frac{\frac{1}{2\tau} - L\sqrt{m-1}}{L^2(m-1)+\frac{1}{\tau^2}} \cdot \frac{1}{3D_{\mathcal{X}}^2}, \quad B = \frac{\frac{1}{2\tau} - L\sqrt{m-1}}{L^2(m-1)+\frac{1}{\tau^2}} \cdot \frac{2}{\sqrt{3}\tau D_{\mathcal{X}}}, \quad C = \frac{1}{2\tau}, \quad \text{and} \quad \xi_k = \|\varepsilon^k\|$$

yields the convergence result.

Appendix B. Proof of Lemmas in the Main Theorems

To prove the two steps mentioned in the sketch at the beginning of Appendix A, the most important result is the following first-order optimality condition (FOC). The proof, which will be postponed to Appendix C.2, is similar to Proposition 8.7 in (Santambrogio, 2015) for $h(x) = x\log x$ and $W = 0$.

Lemma 9 (FOC). Consider the Wasserstein proximal gradient scheme

$$\rho_\tau = \operatorname*{argmin}_{\mu\in\mathscr{P}_2^r(\mathcal{X})} \int_{\mathcal{X}} U(x)\,\mathrm{d}\mu + \int_{\mathcal{X}} h(\mu) + \iint_{\mathcal{X}\times\mathcal{X}} W(x,y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho,\mu).$$

Under Assumption A, ρ_τ satisfies

$$T_{\rho_\tau}^{\rho} - \mathrm{Id} = \tau\bigg(\nabla U + \nabla h'(\rho_\tau) + \nabla\int_{\mathcal{X}} W(\cdot,y)\,\mathrm{d}\rho_\tau(y) + \nabla\int_{\mathcal{X}} W(y,\cdot)\,\mathrm{d}\rho_\tau(y)\bigg)$$

For simplicity, in the following proof, define

$$\mathcal{H}(\rho) = \sum_{j=1}^m \mathcal{H}_j(\rho_j) \quad \text{and} \quad \mathcal{W}(\rho) = \sum_{j=1}^m \mathcal{W}_j(\rho_j)$$

for $\rho = (\rho_1,\dots,\rho_m)$. Also, let

$$\nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j) = \nabla\int_{\mathcal{X}_j} W_j(\cdot,y)\,\mathrm{d}\rho_j(y) + \nabla\int_{\mathcal{X}_j} W_j(y,\cdot)\,\mathrm{d}\rho_j(y)$$

be the Wasserstein gradient of $\mathcal{W}_j$ at ρ_j with respect to W_2 metric. Note that this Wasserstein gradient is a function on $\mathcal{X}_j$.

B.1 Proof of Lemma 1

For simplicity, define

$$V_j^k = \int_{\mathcal{X}_{-j}} V(x_j,x_{-j})\,\mathrm{d}\rho_{-j}^k(x_{-j}), \quad \text{and} \quad V_{-j} = \int_{\mathcal{X}_j} V(x_j,x_{-j})\,\mathrm{d}\rho_j^k(x_j).$$

Step 1. First, let us show

$$\sum_{j=1}^m \mathcal{V}(\rho_j^{k+1}\otimes\rho_{-j}^k) \leqslant m\mathcal{V}(\rho^k) + \mathcal{H}(\rho^k) + \mathcal{W}(\rho^k) - \mathcal{H}(\rho^{k+1}) - \mathcal{W}(\rho^{k+1}) - \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho^{k+1},\rho^k). \tag{21}$$

In fact, by definition of ρ_j^{k+1}, we have

$$\mathcal{V}(\rho_j^{k+1}\otimes\rho_{-j}^k) + \mathcal{H}_j(\rho_j^{k+1}) + \mathcal{W}_j(\rho_j^{k+1}) + \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho_j^{k+1},\rho_j^k) \leqslant \mathcal{V}(\rho^k) + \mathcal{H}_j(\rho_j^k) + \mathcal{W}_j(\rho_j^k).$$

Summing from $j=1$ to m yields equation (21).

Step 2. Next, we will show that

$$\sum_{j=1}^m \mathcal{V}(\rho_j^{k+1}\otimes\rho_{-j}^k) \geqslant m\mathcal{V}(\rho^{k+1}) + (m-1)\sum_{j=1}^m \int_{\mathcal{X}_j} \big\langle \nabla V_j^{k+1}, T_{\rho_j^{k+1}}^{\rho_j^k} - \mathrm{Id}\big\rangle\,\mathrm{d}\rho_j^{k+1}. \tag{22}$$

To prove this, by convexity of $\mathcal{V}$ we have

$$\begin{aligned}\mathcal{V}(\rho_j^{k+1}\otimes\rho_{-j}^k) &\overset{\text{(i)}}{\geqslant} \mathcal{V}(\rho_j^{k+1}\otimes\rho_{-j}^{k+1}) + \int_{\mathcal{X}_{-j}}\left\langle \nabla\frac{\delta\mathcal{V}(\rho_j^{k+1}\otimes\cdot)}{\delta\rho_{-j}}(\rho_{-j}^{k+1}), T^{\rho_{-j}^k}_{\rho_{-j}^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_{-j}^{k+1}\\ &\overset{\text{(ii)}}{=} \mathcal{V}(\rho^{k+1}) + \sum_{i\neq j}\int_{\mathcal{X}_{-j}}\left\langle \nabla_i V_{-j}^{k+1}, T^{\rho_i^k}_{\rho_i^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_{-j}^{k+1}\\ &= \mathcal{V}(\rho^{k+1}) + \sum_{i\neq j}\int_{\mathcal{X}_i}\left\langle \nabla V_i^{k+1}, T^{\rho_i^k}_{\rho_i^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_i^{k+1}\end{aligned}$$

Here, (i) is by Equation (12); (ii) is by

$$\nabla\frac{\delta\mathcal{V}(\rho_j^{k+1}\otimes\cdot)}{\delta\rho_{-j}}(\rho_{-j}^{k+1}) = \nabla V_{-j}^{k+1} = \big(\nabla_1 V_{-j}^{k+1},\dots,\nabla_{j-1}V_{-j}^{k+1},\nabla_{j+1}V_{-j}^{k+1},\dots,\nabla_m V_{-j}^{k+1}\big).$$

Summing the inequality above from $j=1$ to m yields equation (22).

Step 3. Let us show

$$\begin{aligned}&\sum_{j=1}^m\int_{\mathcal{X}_j}\left\langle \nabla V_j^{k+1}, T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1}\\ &\geqslant \big(\tau^{-1} - L\sqrt{m-1}\big)\,\mathrm{W}_2^2(\rho^{k+1},\rho^k) - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big) - \Big(\mathcal{W}_j(\rho_j^k) - \mathcal{W}_j(\rho_j^{k+1})\Big).\end{aligned} \tag{23}$$

By Lemma 9, we have $T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id} = \tau\nabla V_j^k + \tau\nabla h_j'(\rho_j^{k+1}) + \tau\nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^{k+1})$. This implies

$$\begin{aligned}&\sum_{j=1}^m\int_{\mathcal{X}_j}\left\langle \nabla V_j^{k+1}, T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1}\\ &= \sum_{j=1}^m\int_{\mathcal{X}_j}\left\langle \frac{T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}}{\tau} - \big(\nabla V_j^k + \nabla h_j'(\rho_j^{k+1}) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^{k+1})\big), T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle\\ &\qquad + \left\langle \nabla V_j^{k+1}, T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1}\\ &= \sum_{j=1}^m\Bigg[\int_{\mathcal{X}_j}\left\langle \nabla V_j^{k+1} - \nabla V_j^k, T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1} - \int_{\mathcal{X}_j}\left\langle \nabla h_j'(\rho_j^{k+1}), T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1}\\ &\qquad - \int_{\mathcal{X}_j}\left\langle \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^{k+1}), T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1} + \frac{1}{\tau}\,\mathrm{W}_2^2(\rho_j^{k+1},\rho_j^k)\Bigg]\\ &\overset{\text{(i)}}{\geqslant} \sum_{j=1}^m\Bigg[\int_{\mathcal{X}_j}\left\langle \nabla V_j^{k+1} - \nabla V_j^k, T^{\rho_j^k}_{\rho_j^{k+1}} - \mathrm{Id}\right\rangle \mathrm{d}\rho_j^{k+1} - \Big(\mathcal{H}_j(\rho_j^k) - \mathcal{H}_j(\rho_j^{k+1})\Big)\\ &\qquad - \Big(\mathcal{W}_j(\rho_j^k) - \mathcal{W}_j(\rho_j^{k+1})\Big)\Bigg] + \frac{1}{\tau}\,\mathrm{W}_2^2(\rho^{k+1},\rho^k)\\ &\overset{\text{(ii)}}{\geqslant} -\sum_{j=1}^m \mathrm{W}_2(q_j^{k+1},q_j^k)\cdot\left\|\nabla V_j^{k+1} - \nabla V_j^k\right\|_{L^2(\rho_j^{k+1};\mathcal{X}_j)} - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big)\end{aligned}$$

$$
\begin{aligned}
&\qquad - \Big(\mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1})\Big) + \frac{1}{\tau}\,\mathrm{W}_2^2(\rho^{k+1},\rho^k)\\
&\overset{\text{(iii)}}{\geqslant} -\sum_{j=1}^m \mathrm{W}_2(\rho_j^{k+1},\rho_j^k)\cdot L\,\mathrm{W}_2(\rho_{-j}^{k+1},\rho_{-j}^k) - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big)\\
&\qquad - \Big(\mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1})\Big) + \frac{1}{\tau}\,\mathrm{W}_2^2(\rho^{k+1},\rho^k)\\
&\overset{\text{(iv)}}{\geqslant} -L\sum_{j=1}^m\left[\frac{\sqrt{m-1}}{2}\,\mathrm{W}_2^2(\rho_j^{k+1},\rho_j^k) + \frac{1}{2\sqrt{m-1}}\,\mathrm{W}_2^2(\rho_{-j}^{k+1},\rho_{-j}^k)\right]\\
&\qquad - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big) - \Big(\mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1})\Big) + \frac{1}{\tau}\,\mathrm{W}_2^2(\rho^{k+1},\rho^k)\\
&= \big(\tau^{-1} - L\sqrt{m-1}\big)\,\mathrm{W}_2^2(\rho^{k+1},\rho^k) - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big) - \Big(\mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1})\Big).
\end{aligned}
$$

Here, (i) is due to the fact that both $\mathcal{H}_j$ and $\mathcal{W}_j$ are convex along geodesics; (ii) is by Cauchy–Schwarz inequality; (iii) is by Lemma 12; (iv) is by AM-GM inequality.

Step 4. Finally, let us prove the statement. By Equations (21), (22), and (23), we have

$$
\begin{aligned}
&m\mathcal{V}(\rho^k) + \mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1}) + \mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1}) - \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho_j^{k+1},\rho_j^k) \geqslant \sum_{j=1}^m \mathcal{V}(\rho_{-j}^{k+1}\otimes\rho_j^k)\\
&\geqslant m\mathcal{V}(\rho^{k+1}) + (m-1)\sum_{j=1}^m\int_{\mathcal{X}_j}\big\langle\nabla V_j^{k+1}, T_{\rho_j^{k+1}}^{\rho_j^k} - \mathrm{Id}\big\rangle\,\mathrm{d}\rho_j^{k+1}\\
&\geqslant m\mathcal{V}(\rho^{k+1}) + (m-1)\Big[\big(\tau^{-1} - L\sqrt{m-1}\big)\,\mathrm{W}_2^2(\rho^{k+1},\rho^k) - \Big(\mathcal{H}(\rho^k) - \mathcal{H}(\rho^{k+1})\Big)\\
&\qquad\qquad - \Big(\mathcal{W}(\rho^k) - \mathcal{W}(\rho^{k+1})\Big)\Big].
\end{aligned}
$$

This implies

$$
\mathcal{F}(\rho^k) - \mathcal{F}(\rho^{k+1}) \geqslant \frac{1}{m}\left[\frac{1}{2\tau} + (m-1)\Big(\frac{1}{\tau} - L\sqrt{m-1}\Big)\right]\mathrm{W}_2^2(\rho^{k+1},\rho^k).
$$

Substituting k with $k-1$ yields the result.

B.2 Proof of Lemma 2

To prove the result, we need the following lemma.

Lemma 10. Under Assumptions A and B, for both the parallel update scheme and the sequential update scheme, we have

$$
\sum_{j=1}^m\int_{\mathcal{X}_j}\big\|\nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)\big\|^2\,\mathrm{d}\rho_j^k \leqslant \Big(2L^2(m-1) + \frac{2}{\tau^2}\Big)\,\mathrm{W}_2^2(\rho^k,\rho^{k-1}).
$$

Then, by convexity of $\mathcal{F}$, we have

$$
\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant -\int_{\mathcal{X}}\Big\langle\nabla\frac{\delta\mathcal{F}}{\delta\rho}(\rho^k), T_{\rho^k}^{\rho^*} - \mathrm{Id}\Big\rangle\,\mathrm{d}\rho^k
$$

$$
\begin{aligned}
&= -\sum_{j=1}^{m} \int_{\mathcal{X}} \left\langle \nabla_j V + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k), T_{\rho_j^k}^{\rho_j^*} - \mathrm{Id} \right\rangle \mathrm{d}\rho^k \\
&= -\sum_{j=1}^{m} \int_{\mathcal{X}_j} \left\langle \nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k), T_{\rho_j^k}^{\rho_j^*} - \mathrm{Id} \right\rangle \mathrm{d}\rho_j^k \\
&\overset{\text{(i)}}{\leqslant} \sum_{j=1}^{m} \sqrt{\int_{\mathcal{X}_j} \left\| \nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k) \right\|^2 \mathrm{d}\rho_j^k} \cdot \mathrm{W}_2(\rho_j^k, \rho_j^*) \\
&\overset{\text{(ii)}}{\leqslant} \sqrt{\sum_{j=1}^{m} \int_{\mathcal{X}_j} \left\| \nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k) \right\|^2 \mathrm{d}\rho_j^k} \cdot \mathrm{W}_2(\rho^k, \rho^*) \\
&\overset{\text{(iii)}}{\leqslant} \sqrt{2L^2(m-1) + \frac{2}{\tau^2}}\, \mathrm{W}_2(\rho^k, \rho^{k-1})\, \mathrm{W}_2(\rho^k, \rho^*). \qquad (24)
\end{aligned}
$$

Here, both (i) and (ii) are due to Cauchy–Schwarz inequality, and (iii) is by Lemma 10. By Assumption D (λ-QG condition), we have

$$
\frac{\lambda}{2} \mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \mathcal{F}(\rho^t) - \mathcal{F}(\rho^*) \leqslant \sqrt{2L^2(m-1) + \frac{2}{\tau^2}}\, \mathrm{W}_2(\rho^k, \rho^{k-1})\, \mathrm{W}_2(\rho^k, \rho^*).
$$

This implies

$$
\mathrm{W}_2(\rho^k, \rho^*) \leqslant \frac{2}{\lambda} \sqrt{2L^2(m-1) + \frac{2}{\tau^2}}\, \mathrm{W}_2(\rho^k, \rho^{k-1}).
$$

Substituting the term $\mathrm{W}_2(\rho^k, \rho^*)$ in (24) with the above inequality yields

$$
\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{2}{\lambda} \Big(2L^2(m-1) + \frac{2}{\tau^2} \Big) \mathrm{W}_2^2(\rho^k, \rho^{k-1}).
$$

B.3 Proof of Lemma 3

Proof By the definition of ρ_j^k, we have

$$
\begin{aligned}
&\mathcal{V}(\rho_1^{k-1}, \ldots, \rho_j^{k-1}, \rho_{j+1}^k, \ldots, \rho_m^k) + \mathcal{H}_j(\rho_j^{k-1}) + \mathcal{W}_j(\rho_j^{k-1}) \\
\geqslant\ &\mathcal{V}(\rho_1^{k-1}, \ldots, \rho_{j-1}^{k-1}, \rho_j^k, \ldots, \rho_m^k) + \mathcal{H}_j(\rho_j^k) + \mathcal{W}_j(\rho_j^k) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho_j^{k-1}, \rho_j^k).
\end{aligned}
$$

This implies

$$
\begin{aligned}
\mathcal{F}(\rho^{k-1}) - \mathcal{F}(\rho^k) &= \sum_{j=1}^{m} \Big[\mathcal{V}(\rho_1^{k-1}, \ldots, \rho_j^{k-1}, \rho_j^{k+1}, \ldots, \rho_m^k) + \mathcal{H}_j(\rho_j^{k-1}) + \mathcal{W}_j(\rho_j^{k-1}) \\
&\qquad - \mathcal{V}(\rho_1^{k-1} \ldots, \rho_{j-1}^{k-1}, \rho_j^k, \ldots, \rho_m^k) - \mathcal{H}_j(\rho_j^k) - \mathcal{W}_j(\rho_j^k) \Big] \\
&\geqslant \sum_{j=1}^{m} \frac{1}{2\tau} \mathrm{W}_2^2(\rho_j^k, \rho_j^{k-1}) \\
&= \frac{1}{2\tau} \mathrm{W}_2^2(\rho^k, \rho^{k-1}).
\end{aligned}
$$

■

B.4 Proof of Lemma 4

Proof Let $Y_j = \sum_{l=0}^{M-1} I\{j_l = j\}$ and $A_j = \{Y_j = 0\} \cup \{Y_j > T\}$ be a event. Then

$$\mathbb{P}\big(A_1 \bigcup \cdots \bigcup A_m\big) \leqslant \sum_{j=1}^{m} \mathbb{P}(A_j) = m\mathbb{P}(Y_1 = 0) + m\mathbb{P}(Y_1 > T).$$

Take $M = Rm\log m$ with some R to be decided later. Note that

$$\mathbb{P}(Y_1 = 0) = \Big(1 - \frac{1}{m}\Big)^M = \left[\frac{1}{\big(1 + \frac{1}{m-1}\big)^m}\right]^{\frac{M}{m}} \leqslant e^{-\frac{M}{m}} = e^{-R\log m} = m^{-R}.$$

Applying Chernoff's inequality to Bernoulli distribution (Theorem 2.3.1 in(Vershynin, 2018)) yields

$$\mathbb{P}\Big(Y_1 > eR\log m\Big) \leqslant e^{-R\log m} = m^{-R}.$$

Therefore, we have

$$\mathbb{P}\big(A_1 \bigcup \cdots \bigcup A_m\big) \leqslant \sum_{j=1}^{m} \mathbb{P}(A_j) \leqslant 2m \cdot m^{-R} = 2m^{1-R}.$$

By taking

$$R = 1 - \frac{\log\delta}{\log m}$$

we have $2m^{1-R} = 2\delta$. ■

B.5 Proof of Lemma 5

Proof By the definition of $\rho_{j_l}^{k,l+1}$, we have

$$\begin{aligned}\mathcal{V}(\rho_{j_l}^{k,l+1}, \rho_{-j_l}^{k,l}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l+1}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l+1}) + \frac{1}{2\tau}\mathrm{W}_2^2(\rho_{j_l}^{k,l+1}, \rho_{j_l}^{k,l}) \\ \leqslant \mathcal{V}(\rho_{j_l}^{k,l}, \rho_{-j_l}^{k,l}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l}).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathcal{F}(\rho^k) - \mathcal{F}(\rho^{k+1}) = \mathcal{F}(\rho^{k,0}) - \mathcal{F}(\rho^{k,M}) = \sum_{l=0}^{M-1}\big[\mathcal{F}(\rho^{k,l}) - \mathcal{F}(\rho^{k,l+1})\big] \\ &= \sum_{l=0}^{M-1}\big[\mathcal{V}(\rho_{j_l}^{k,l}, \rho_{-j_l}^{k,l}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l})\big] - \big[\mathcal{V}(\rho_{j_l}^{k,l+1}, \rho_{-j_l}^{k,l+1}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l+1}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l+1})\big] \\ &\overset{\text{(i)}}{=} \sum_{l=0}^{M-1}\big[\mathcal{V}(\rho_{j_l}^{k,l}, \rho_{-j_l}^{k,l}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l})\big] - \big[\mathcal{V}(\rho_{j_l}^{k,l+1}, \rho_{-j_l}^{k,l}) + \mathcal{H}_{j_l}(\rho_{j_l}^{k,l+1}) + \mathcal{W}_{j_l}(\rho_{j_l}^{k,l+1})\big]\end{aligned}$$

$$\geqslant \sum_{l=0}^{M-1} \frac{1}{2\tau} \mathrm{W}_2^2(\rho_{j_l}^{k,l}, \rho_{j_l}^{k,l+1}) \overset{\text{(ii)}}{=} \frac{1}{2\tau} \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k,l}, \rho^{k,l+1}).$$

Here, both (i) and (ii) are due to $\rho_{-j_l}^{k,l+1} = \rho_{-j_l}^{k,l}$. ■

B.6 Proof of Lemma 6

Proof Similar to the argument in Lemma 2, we have

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \sqrt{\sum_{j=1}^{m} \int_{\mathcal{X}_j} \left\| \nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k) \right\|^2 \mathrm{d}\rho_j^k} \cdot \mathrm{W}_2(\rho^k, \rho^*).$$

To bound the norm of the blockwise Wasserstein gradient, we need the following lemma.

Lemma 11. For any $k \in \mathbb{Z}_+$, let $M \in \mathbb{Z}_+$ be an integer such that all coordinates are updated at least once but at most T times through M updates in the k-th iteration. Under Assumptions A and B, we have

$$\sum_{j=1}^{m} \int_{\mathcal{X}_j} \left\| \nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^k) \right\|^2 \mathrm{d}\rho_j^k \leqslant \Big(2L^2 mT + \frac{2}{\tau^2}\Big) \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}).$$

With the above lemma, we have

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \sqrt{\Big(2L^2 mT + \frac{2}{\tau^2}\Big) \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1})} \cdot \mathrm{W}_2(\rho^k, \rho^*). \tag{25}$$

By Assumption D (λ-QG), we have

$$\frac{\lambda}{2} \mathrm{W}_2^2(\rho^k, \rho^*) \leqslant \mathcal{F}(\rho^k) - \mathcal{F}(\rho^*).$$

Combining the above two pieces yields

$$\frac{\lambda}{2} \mathrm{W}_2(\rho^k, \rho^*) \leqslant \sqrt{\Big(2L^2 mT + \frac{2}{\tau^2}\Big) \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1})},$$

which implies

$$\mathcal{F}(\rho^k) - \mathcal{F}(\rho^*) \leqslant \frac{4}{\lambda}\Big(L^2 mT + \frac{1}{\tau^2}\Big) \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}).$$

■

B.7 Proof of Lemma 7

Step 1. By convexity of $\mathcal{F}$, we have

$$\begin{aligned}
&\big[\mathcal{V}(\widetilde{\rho}_j^k\otimes\widetilde{\rho}_{-j}^k)+\mathcal{H}_j(\widetilde{\rho}_j^k)+\mathcal{W}_j(\widetilde{\rho}_j^k)\big]-\big[\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\widetilde{\rho}_{-j}^k)+\mathcal{H}_j(\widetilde{\rho}_j^{k+1})+\mathcal{W}_j(\widetilde{\rho}_j^{k+1})\big]\\
&\geqslant\int_{\mathcal{X}_j}\Big\langle\nabla V_j^k+\nabla h_j'(\widetilde{\rho}_j^{k+1})+\nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^{k+1}),T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}\\
&=\frac{1}{\tau}\int_{\mathcal{X}_j}\Big\langle T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}-\eta_j^{k+1},T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}\\
&=\frac{1}{\tau}\,\mathrm{W}_2^2(\widetilde{\rho}_j^{k+1},\widetilde{\rho}_j^k)-\frac{1}{\tau}\int_{\mathcal{X}_j}\Big\langle\eta_j^{k+1},T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}\\
&\geqslant\frac{1}{\tau}\,\mathrm{W}_2^2(\widetilde{\rho}_j^{k+1},\widetilde{\rho}_j^k)-\frac{1}{\tau}\bigg(\frac{1}{2}\int_{\mathcal{X}_j}\|\eta_j^{k+1}\|^2\,\mathrm{d}\widetilde{\rho}_j^{k+1}+\frac{1}{2}\,\mathrm{W}_2^2(\widetilde{\rho}_j^{k+1},\widetilde{\rho}_j^k)\bigg)\\
&\geqslant\frac{1}{2\tau}\,\mathrm{W}_2^2(\widetilde{\rho}_j^{k+1},\widetilde{\rho}_j^k)-\frac{(\varepsilon_j^{k+1})^2}{2\tau}.
\end{aligned}$$

Summing j from 1 to m yields

$$\begin{aligned}
\sum_{j=1}^m\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\widetilde{\rho}_{-j}^k)\leqslant{}&m\mathcal{V}(\widetilde{\rho}^k)+\mathcal{H}(\widetilde{\rho}^k)+\mathcal{W}(\widetilde{\rho}^k)\\
&-\mathcal{H}(\widetilde{\rho}^{k+1})-\mathcal{W}(\widetilde{\rho}^k)-\frac{1}{2\tau}\,\mathrm{W}_2^2(\widetilde{\rho}^{k+1},\widetilde{\rho}^k)+\frac{\|\varepsilon^{k+1}\|^2}{2\tau}.
\end{aligned}$$

Step 2. By convexity of $\mathcal{V}$, we have

$$\begin{aligned}
\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\widetilde{\rho}_{-j}^k)&\geqslant\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\widetilde{\rho}_{-j}^{k+1})+\int_{\mathcal{X}_{-j}}\Big\langle\nabla\frac{\delta\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\cdot)}{\delta\rho_{-j}}(\widetilde{\rho}_{-j}^{k+1}),T_{\widetilde{\rho}_{-j}^{k+1}}^{\widetilde{\rho}_{-j}^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_{-j}^{k+1}\\
&=\mathcal{V}(\widetilde{\rho}^{k+1})+\sum_{i\neq j}\int_{\mathcal{X}_i}\Big\langle\nabla V_i^{k+1},T_{\widetilde{\rho}_i^{k+1}}^{\widetilde{\rho}_i^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_i^{k+1}.
\end{aligned}$$

Summing j from 1 to m yields

$$\sum_{j=1}^m\mathcal{V}(\widetilde{\rho}_j^{k+1}\otimes\widetilde{\rho}_{-j}^k)\geqslant m\mathcal{V}(\widetilde{\rho}^{k+1})+(m-1)\sum_{j=1}^m\int_{\mathcal{X}_j}\Big\langle\nabla V_j^{k+1},T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}.$$

Step 3. We have

$$\begin{aligned}
&\sum_{j=1}^m\int_{\mathcal{X}_j}\Big\langle\nabla V_j^{k+1},T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}\\
&=\sum_{j=1}^m\int_{\mathcal{X}_j}\Big\langle\frac{T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}-\eta_j^{k+1}}{\tau}-\big(\nabla V_j^k+\nabla h_j'(\widetilde{\rho}_j^{k+1})+\nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^{k+1})\big),T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\\
&\qquad+\Big\langle\nabla V_j^{k+1},T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k}-\mathrm{Id}\Big\rangle\,\mathrm{d}\widetilde{\rho}_j^{k+1}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^m \Bigg[\int_{\mathcal{X}_j} \Big\langle \nabla V_j^{k+1} - \nabla V_j^k, T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} - \int_{\mathcal{X}_j} \Big\langle \nabla h_j'(\widetilde{\rho}_j^{k+1}), T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} \\
&\qquad - \int_{\mathcal{X}_j} \Big\langle \nabla_{\mathrm{W}_2} \mathcal{W}_j(\widetilde{\rho}_j^{k+1}), T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} + \frac{1}{\tau} \mathrm{W}_2^2(\widetilde{\rho}_j^{k+1}, \widetilde{\rho}_j^k) \\
&\qquad - \frac{1}{\tau} \int_{\mathcal{X}_j} \Big\langle \eta_j^{k+1}, T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} \Bigg] \\
&\geqslant \sum_{j=1}^m \Bigg[\int_{\mathcal{X}_j} \Big\langle \nabla V_j^{k+1} - \nabla V_j^k, T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} - \int_{\mathcal{X}_j} \Big\langle \nabla h_j'(\widetilde{\rho}_j^{k+1}), T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} \\
&\qquad - \int_{\mathcal{X}_j} \Big\langle \nabla_{\mathrm{W}_2} \mathcal{W}_j(\widetilde{\rho}_j^{k+1}), T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} + \frac{1}{2\tau} \mathrm{W}_2^2(\widetilde{\rho}_j^{k+1}, \widetilde{\rho}_j^k) - \frac{(\varepsilon_j^{k+1})^2}{2\tau} \Bigg] \\
&\geqslant - \sum_{j=1}^m \mathrm{W}_2(\widetilde{\rho}_j^{k+1}, \widetilde{\rho}_j^k) \cdot \Big\| \nabla V_j^{k+1} - \nabla V_j^k \Big\|_{L^2(\widetilde{\rho}_j^{k+1}; \mathcal{X}_j)} - \big[\mathcal{H}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) \big] \\
&\qquad - \big[\mathcal{W}(\widetilde{\rho}^k) - \mathcal{W}(\widetilde{\rho}^{k+1}) \big] + \frac{1}{2\tau} \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \frac{\|\varepsilon^{k+1}\|^2}{2\tau} \\
&\geqslant - \sum_{j=1}^m \mathrm{W}_2(\widetilde{\rho}_j^{k+1}, \widetilde{\rho}_j^k) \cdot L\, \mathrm{W}_2(\widetilde{\rho}_{-j}^{k+1}, \widetilde{\rho}_{-j}^k) - \big[\mathcal{H}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) \big] \\
&\qquad - \big[\mathcal{W}(\widetilde{\rho}^k) - \mathcal{W}(\widetilde{\rho}^{k+1}) \big] + \frac{1}{2\tau} \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \frac{\|\varepsilon^{k+1}\|^2}{2\tau} \\
&\geqslant -L \sum_{j=1}^m \left[\frac{\sqrt{m-1}}{2} \mathrm{W}_2^2(\widetilde{\rho}_j^{k+1}, \widetilde{\rho}_j^k) + \frac{1}{2\sqrt{m-1}} \mathrm{W}_2^2(\widetilde{\rho}_{-j}^{k+1}, \widetilde{\rho}_{-j}^k) \right] - \big[\mathcal{H}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) \big] \\
&\qquad - \big[\mathcal{W}(\widetilde{\rho}^k) - \mathcal{W}(\widetilde{\rho}^{k+1}) \big] + \frac{1}{2\tau} \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \frac{\|\varepsilon^{k+1}\|^2}{2\tau} \\
&= \Big(\frac{1}{2\tau} - L\sqrt{m-1} \Big) \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \big[\mathcal{H}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) \big] - \big[\mathcal{W}(\widetilde{\rho}^k) - \mathcal{W}(\widetilde{\rho}^{k+1}) \big] - \frac{\|\varepsilon^{k+1}\|^2}{2\tau}.
\end{aligned}
$$

Step 4. Combining all pieces above yields

$$
\begin{aligned}
&m\mathcal{V}(\widetilde{\rho}^k) + \mathcal{H}(\widetilde{\rho}^k) + \mathcal{W}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) - \mathcal{W}(\widetilde{\rho}^k) - \frac{1}{2\tau} \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) + \frac{\|\varepsilon^{k+1}\|^2}{2\tau} \\
&\geqslant \sum_{j=1}^m \mathcal{V}(\widetilde{\rho}_j^{k+1} \otimes \widetilde{\rho}_{-j}^k) \geqslant m\mathcal{V}(\widetilde{\rho}^{k+1}) + (m-1) \sum_{j=1}^m \int_{\mathcal{X}_j} \Big\langle \nabla V_j^{k+1}, T_{\widetilde{\rho}_j^{k+1}}^{\widetilde{\rho}_j^k} - \mathrm{Id} \Big\rangle \mathrm{d}\widetilde{\rho}_j^{k+1} \\
&\geqslant m\mathcal{V}(\widetilde{\rho}^{k+1}) + (m-1) \Bigg[\Big(\frac{1}{2\tau} - L\sqrt{m-1} \Big) \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \big[\mathcal{H}(\widetilde{\rho}^k) - \mathcal{H}(\widetilde{\rho}^{k+1}) \big] \\
&\qquad - \big[\mathcal{W}(\widetilde{\rho}^k) - \mathcal{W}(\widetilde{\rho}^{k+1}) \big] - \frac{\|\varepsilon^{k+1}\|^2}{2\tau} \Bigg].
\end{aligned}
$$

This implies

$$
\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\widetilde{\rho}^{k+1}) \geqslant \Big(\frac{1}{2\tau} - L\sqrt{m-1} \Big) \mathrm{W}_2^2(\widetilde{\rho}^{k+1}, \widetilde{\rho}^k) - \frac{\|\varepsilon^{k+1}\|^2}{2\tau}.
$$

B.8 Proof of Lemma 8

Step 1. We first prove that

$$\sum_{j=1}^m \int_{\mathcal{X}_j} \left\|\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k)\right\|^2 \mathrm{d}\widetilde{\rho}_j^k \leqslant \Big(3L^2(m-1) + \frac{3}{\tau^2}\Big)\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}.$$

Just note that

$$\begin{aligned}
&\int_{\mathcal{X}_j} \left\|\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k)\right\|^2 \mathrm{d}\widetilde{\rho}_j^k \\
&= \int_{\mathcal{X}_j} \Big\|\nabla V_j^k - \nabla V_j^{k-1} + \frac{1}{\tau}\Big(T_{\widetilde{\rho}_j^k}^{\widetilde{\rho}_j^{k-1}} - \mathrm{Id} - \eta_j^k\Big)\Big\|^2 \mathrm{d}\widetilde{\rho}_j^k \\
&\leqslant 3\int_{\mathcal{X}_j} \left\|\nabla V_j^k - \nabla V_j^{k-1}\right\|^2 \mathrm{d}\widetilde{\rho}_j^k + \frac{3}{\tau^2}\,\mathrm{W}_2^2(\widetilde{\rho}_j^k, \widetilde{\rho}_j^{k-1}) + \frac{3(\varepsilon_j^k)^2}{\tau^2} \\
&\leqslant 3L^2\,\mathrm{W}_2^2(\widetilde{\rho}_{-j}^{k-1}, \widetilde{\rho}_{-j}^k) + \frac{3}{\tau^2}\,\mathrm{W}_2^2(\widetilde{\rho}_j^k, \widetilde{\rho}_j^{k-1}) + \frac{3(\varepsilon_j^k)^2}{\tau^2}.
\end{aligned}$$

Summing j from 1 to m yields

$$\sum_{j=1}^m \int_{\mathcal{X}_j} \left\|\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k)\right\|^2 \mathrm{d}\widetilde{\rho}_j^k \leqslant \Big(3L^2(m-1) + \frac{3}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}.$$

Step 2. By convexity of $\mathcal{F}$, we have

$$\begin{aligned}
\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) &\leqslant -\int_{\mathcal{X}} \Big\langle \nabla\frac{\delta\mathcal{F}}{\delta\rho}(\widetilde{\rho}^k), T_{\widetilde{\rho}^k}^{\rho^*} - \mathrm{Id}\Big\rangle \mathrm{d}\widetilde{\rho}^k \\
&= -\sum_{j=1}^m \int_{\mathcal{X}} \Big\langle \nabla_j V + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k), T_{\widetilde{\rho}_j^k}^{\rho_j^*} - \mathrm{Id}\Big\rangle \mathrm{d}\widetilde{\rho}^k \\
&= -\sum_{j=1}^m \int_{\mathcal{X}_j} \Big\langle \nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k), T_{\widetilde{\rho}_j^k}^{\rho_j^*} - \mathrm{Id}\Big\rangle \mathrm{d}\widetilde{\rho}_j^k \\
&\leqslant \sum_{j=1}^m \sqrt{\int_{\mathcal{X}_j} \left\|\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k)\right\|^2 \mathrm{d}\widetilde{\rho}_j^k} \cdot \mathrm{W}_2(\widetilde{\rho}_j^k, \rho_j^*) \\
&\leqslant \sqrt{\sum_{j=1}^m \int_{\mathcal{X}_j} \left\|\nabla V_j^k + \nabla h_j'(\widetilde{\rho}_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\widetilde{\rho}_j^k)\right\|^2 \mathrm{d}\widetilde{\rho}_j^k} \cdot \mathrm{W}_2(\widetilde{\rho}^k, \rho^*) \\
&\leqslant \sqrt{\Big(3L^2(m-1) + \frac{3}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}} \cdot \mathrm{W}_2(\widetilde{\rho}^k, \rho^*). \qquad (26)
\end{aligned}$$

By λ-QG condition, we have

$$\frac{\lambda}{2}\,\mathrm{W}_2^2(\widetilde{\rho}^k, \rho^*) \leqslant \mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) \leqslant \sqrt{\Big(3L^2(m-1) + \frac{3}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}} \cdot \mathrm{W}_2(\widetilde{\rho}^k, \rho^*).$$

So, we have

$$\mathrm{W}_2(\widetilde{\rho}^k, \rho^*) \leqslant \frac{2}{\lambda}\sqrt{\Big(3L^2(m-1)+\frac{3}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{3\|\varepsilon^k\|^2}{\tau^2}},$$

indicating that

$$\mathcal{F}(\widetilde{\rho}^k) - \mathcal{F}(\rho^*) \leqslant \frac{6}{\lambda}\Big[\Big(L^2(m-1)+\frac{1}{\tau^2}\Big)\,\mathrm{W}_2^2(\widetilde{\rho}^k, \widetilde{\rho}^{k-1}) + \frac{\|\varepsilon^k\|^2}{\tau^2}\Big].$$

B.9 Proof of Lemma 10

The proof for two schemes are slightly different. We will consider these two cases separately.

Parallel update scheme. By Lemma 9, $T_{\rho_j^k}^{\rho_j^{k-1}} - \mathrm{Id} = \tau\nabla V_j^{k-1} + \tau\nabla h_j'(\rho_j^k) + \tau\nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)$. Therefore, we have

$$\begin{aligned}
\int_{\mathcal{X}_j} \big\|\nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)\big\|^2 \,\mathrm{d}\rho_j^k &= \int_{\mathcal{X}_j} \Big\|\nabla V_j^k - \nabla V_j^{k-1} + \frac{1}{\tau}\Big(T_{\rho_j^k}^{\rho_j^{k-1}} - \mathrm{Id}\Big)\Big\|^2 \,\mathrm{d}\rho_j^k \\
&\overset{\text{(i)}}{\leqslant} 2\int_{\mathcal{X}_j} \big\|\nabla V_j^k - \nabla V_j^{k-1}\big\|^2 \,\mathrm{d}\rho_j^k + \frac{2\,\mathrm{W}_2^2(\rho_j^k, \rho_j^{k-1})}{\tau^2} \\
&\overset{\text{(ii)}}{\leqslant} 2L^2\,\mathrm{W}_2^2(\rho_{-j}^{k-1}, \rho_{-j}^k) + \frac{2\,\mathrm{W}_2^2(\rho_j^k, \rho_j^{k-1})}{\tau^2}.
\end{aligned}$$

Here, (i) is by Cauchy–Schwarz inequality, and (ii) is by Lemma 12. Summing the above inequality from $j = 1$ to m yields the desired result.

Sequential update scheme. For simplicity, let

$$\widetilde{V}_j^k = \int_{\mathcal{X}_{-j}} V(x_j, x_{-j})\,\mathrm{d}\rho_1^k \cdots \mathrm{d}\rho_{j-1}^k\,\mathrm{d}\rho_{j+1}^{k-1}\cdots\mathrm{d}\rho_m^{k-1}$$

be a function on $\mathcal{X}_j$. By Lemma 9, we have $T_{\rho_j^k}^{\rho_j^{k-1}} - \mathrm{Id} = \tau\big[\nabla\widetilde{V}_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)\big]$. Therefore, we have

$$\begin{aligned}
&\sum_{j=1}^m \int_{\mathcal{X}_j} \big\|\nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)\big\|^2 \,\mathrm{d}\rho_j^k \\
&= \sum_{j=1}^m \int_{\mathcal{X}_j} \Big\|\nabla V_j^k - \nabla\widetilde{V}_j^k + \frac{1}{\tau}\Big(T_{\rho_j^k}^{\rho_j^{k-1}} - \mathrm{Id}\Big)\Big\|^2 \,\mathrm{d}\rho_j^k \\
&\overset{\text{(i)}}{\leqslant} 2\sum_{j=1}^m \int_{\mathcal{X}_j} \big\|\nabla V_j^k - \nabla\widetilde{V}_j^k\big\|^2 \,\mathrm{d}\rho_j^k + \frac{2}{\tau^2}\sum_{j=1}^m \int_{\mathcal{X}_j} \Big\|T_{\rho_j^k}^{\rho_j^{k-1}} - \mathrm{Id}\Big\|^2 \,\mathrm{d}\rho_j^k \\
&\overset{\text{(ii)}}{\leqslant} 2\sum_{j=1}^m L^2 \sum_{l=j+1}^m \mathrm{W}_2^2(\rho_l^k, \rho_l^{k-1}) + \frac{2}{\tau^2}\,\mathrm{W}_2^2(\rho^k, \rho^{k-1}) \\
&\leqslant \Big[2L^2(m-1)+\frac{2}{\tau^2}\Big]\,\mathrm{W}_2^2(\rho^k, \rho^{k-1}).
\end{aligned}$$

Here, (i) is by Cauchy–Schwarz inequality; (ii) is by Lemma 12.

B.10 Proof of Lemma 11

Proof For any $j \in [m]$ and $k \in \mathbb{N}$, let $I_{j,k} \in [M]$ be the time of the latest update of the j-th coordinate in the k-th iteration. By definition, we know $\rho_j^{k-1,M} = \rho_j^k = \rho_j^{k-1,I_{j,k}}$ for all $j \in [m]$. Since

$$\rho_j^{k-1,I_{j,k}} = \operatorname*{argmin}_{\rho_j \in \mathscr{P}_2^r(\mathcal{X}_j)} \mathcal{V}\big(\rho_j, \rho_{-j}^{k-1,I_{j,k}-1}\big) + \mathcal{H}_j(\rho_j) + \mathcal{W}_j(\rho_j) + \frac{1}{2\tau}\,\mathrm{W}_2^2\big(\rho_j, \rho_j^{k-1,I_{j,k}-1}\big),$$

Lemma 9 implies

$$T_{\rho_j^{k-1,I_{j,k}}}^{\rho_j^{k-1,I_{j,k}-1}} - \mathrm{Id} = \tau\Big(\nabla V_j^{k-1,I_{j,k}-1} + \nabla h_j'\big(\rho_j^{k-1,I_{j,k}}\big) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^{k-1,I_{j,k}})\Big),$$

where we let

$$V_j^{k-1,I_{j,k}-1}(x_j) = \int_{\mathcal{X}_{-j}} V(x_j, x_{-j})\,\mathrm{d}\rho_{-j}^{k-1,I_{j,k}-1}$$

Therefore, we have

$$\begin{aligned}
&\sum_{j=1}^m \int_{\mathcal{X}_j} \big\|\nabla V_j^k + \nabla h_j'(\rho_j^k) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^k)\big\|^2\,\mathrm{d}\rho_j^k \\
&= \sum_{j=1}^m \int_{\mathcal{X}_j} \big\|\nabla V_j^k + \nabla h_j'(\rho_j^{k-1,I_{j,k}}) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^{k-1,I_{j,k}})\big\|^2\,\mathrm{d}\rho_j^k \\
&= \sum_{j=1}^m \int_{\mathcal{X}_j} \Big\|\nabla V_j^k - \nabla V_j^{k-1,I_{j,k}-1} + \frac{1}{\tau}\Big(T_{\rho_j^{k-1,I_{j,k}}}^{\rho_j^{k-1,I_{j,k}-1}} - \mathrm{Id}\Big)\Big\|^2\,\mathrm{d}\rho_j^k \\
&\leqslant 2\sum_{j=1}^m \int_{\mathcal{X}_j} \big\|\nabla V_j^k - \nabla V_j^{k-1,I_{j,k}-1}\big\|^2\,\mathrm{d}\rho_j^k + \frac{2}{\tau^2}\sum_{j=1}^m \int_{\mathcal{X}_j} \Big\|T_{\rho_j^{k-1,I_{j,k}}}^{\rho_j^{k-1,I_{j,k}-1}} - \mathrm{Id}\Big\|^2\,\mathrm{d}\rho_j^k \\
&\overset{\text{(i)}}{\leqslant} 2L^2\sum_{j=1}^m \mathrm{W}_2^2\big(\rho_{-j}^k, \rho_{-j}^{k-1,I_{j,k}-1}\big) + \frac{2}{\tau^2}\sum_{j=1}^m \mathrm{W}_2^2\big(\rho_j^{k-1,I_{j,k}-1}, \rho_j^{k-1,I_{j,k}}\big) \\
&= 2L^2\sum_{j=1}^m \mathrm{W}_2^2\big(\rho_{-j}^{k-1,M}, \rho_{-j}^{k-1,I_{j,k}-1}\big) + \frac{2}{\tau^2}\sum_{j=1}^m \mathrm{W}_2^2\big(\rho_j^{k-1,I_{j,k}-1}, \rho_j^{k-1,I_{j,k}}\big). \qquad (27)
\end{aligned}$$

Here, (i) is by Lemma 12. In the k-th iteration, recall that $j_l \in [m]$ be the coordinate updated in the l-th update. The goal is to bound (27).

Bound of the second term in (27). It is obvious that

$$\sum_{j=1}^m \mathrm{W}_2^2\big(\rho_j^{k-1,I_{j,k}-1}, \rho_j^{k-1,I_{j,k}}\big) \leqslant \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho_{j_l}^{k-1,l}, \rho_{j_l}^{k-1,l+1}) = \sum_{l=0}^{M-1} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l+1}).$$

The first inequality is because each coordinate is updated at least once, and the second equation is because only the j_l-th coordinate is updated in the l-th update.

Bound of the first term in (27). Note that

$$\sum_{j=1}^{m} \mathrm{W}_2^2\left(\rho_{-j}^{k-1,M}, \rho_{-j}^{k-1,I_{j,k}-1}\right) \leqslant \sum_{j=1}^{m} \mathrm{W}_2^2\left(\rho^{k-1,M}, \rho^{k-1,I_{j,k}-1}\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{W}_2^2\left(\rho_i^{k-1,M}, \rho_i^{k-1,I_{j,k}-1}\right).$$

To bound this sum, define the set of iteration

$$S_j^{k-1}(a,b) = \{a \leqslant l \leqslant b : \rho_j^{k-1,l-1} \neq \rho_j^{k-1,l}\}.$$

We know $|S_j^{k-1}(1,M)| \leqslant T$ since each coordinate is updated at most T times. Thus

$$\begin{aligned}
\sum_{i,j=1}^{m} \mathrm{W}_2^2\left(\rho_i^{k-1,M}, \rho_i^{k-1,I_{j,k}-1}\right) &\leqslant \sum_{i,j=1}^{m} \Big(\sum_{l\in S_i^{k-1}(I_{j,k},M)} \mathrm{W}_2(\rho_i^{k-1,l-1}, \rho_i^{k-1,l})\Big)^2 \\
&\leqslant \sum_{i,j=1}^{m} |S_i^{k-1}(I_{j,k},M)| \cdot \sum_{l\in S_i^{k-1}(I_{j,k},M)} \mathrm{W}_2^2(\rho_i^{k-1,l-1}, \rho_i^{k-1,l}) \\
&\leqslant T \sum_{i,j=1}^{m} \sum_{l\in S_i^{k-1}(I_{j,k},M)} \mathrm{W}_2^2(\rho_i^{k-1,l-1}, \rho_i^{k-1,l}) \\
&\leqslant T \sum_{i,j=1}^{m} \sum_{l=1}^{M} \mathrm{W}_2^2(\rho_i^{k-1,l-1}, \rho_i^{k-1,l}) \\
&= mT \sum_{l=1}^{M} \mathrm{W}_2^2(\rho^{k-1,l}, \rho^{k-1,l-1}).
\end{aligned}$$

Combining all pieces above yields the desired result. ∎

Appendix C. Other Technical Lemmas

C.1 Proof of Proposition 8

$$\begin{aligned}
&\mathcal{F}(\rho_1,\ldots,\rho_m) - \mathcal{F}(\rho_1^*,\ldots,\rho_m^*) \geqslant \int_{\mathcal{X}} \Big\langle \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho^*), T_{\rho^*}^{\rho} - \mathrm{Id} \Big\rangle \,\mathrm{d}\rho^* + \frac{\lambda}{2}\,\mathrm{W}_2^2(\rho,\rho^*) \\
&= \int_{\mathcal{X}} \big\langle \nabla V, T_{\rho^*}^{\rho} - \mathrm{Id} \big\rangle \,\mathrm{d}\rho^* + \sum_{j=1}^{m} \int_{\mathcal{X}_j} \big\langle h_j'(\rho_j^*) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^*), T_{\rho_j^*}^{\rho_j} - \mathrm{Id} \big\rangle \,\mathrm{d}\rho_j^* + \frac{\lambda}{2} \sum_{j=1}^{m} \mathrm{W}_2^2(\rho_j,\rho_j^*) \\
&= \sum_{j=1}^{m} \int_{\mathcal{X}} \big\langle \nabla_j V, T_{\rho_j^*}^{\rho_j} - \mathrm{Id} \big\rangle \,\mathrm{d}\rho^* + \sum_{j=1}^{m} \int_{\mathcal{X}_j} \big\langle h_j'(\rho_j^*) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^*), T_{\rho_j^*}^{\rho_j} - \mathrm{Id} \big\rangle \,\mathrm{d}\rho_j^* + \frac{\lambda}{2} \sum_{j=1}^{m} \mathrm{W}_2^2(\rho_j,\rho_j^*) \\
&= \sum_{j=1}^{m} \int_{\mathcal{X}_j} \big\langle \nabla V_j^* + h_j'(\rho_j^*) + \nabla_{\mathrm{W}_2} \mathcal{W}_j(\rho_j^*), T_{\rho_j^*}^{\rho_j} - \mathrm{Id} \big\rangle \,\mathrm{d}\rho_j^* + \frac{\lambda}{2} \sum_{j=1}^{m} \mathrm{W}_2^2(\rho_j,\rho_j^*),
\end{aligned}$$

where we let

$$V_j^* = \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \,\mathrm{d}\rho_{-j}^*$$

for simplicity. Since ρ^* is the stationary point of the optimization problem (1) on the Wasserstein space, we know $\nabla V_j^* + h_j'(\rho_j^*) + \nabla_{\mathrm{W}_2}\mathcal{W}_j(\rho_j^*) = 0$ (e.g., using Lemma 10). This implies

$$\mathcal{F}(\rho_1, \ldots, \rho_m) - \mathcal{F}(\rho_1^*, \ldots, \rho_m^*) \geqslant \frac{\lambda}{2}\sum_{j=1}^m \mathrm{W}_2^2(\rho_j, \rho_j^*).$$

C.2 Proof of Lemma 9 (first-order optimality condition)

Proof Since $\rho \in \mathscr{P}_2^r(\mathcal{X})$, by Brenier's theorem (Theorem 2.12 in (Villani, 2021)), there is a unique optimal map $\nabla\phi$ from ρ_τ to ρ and the Kantorovich potential ϕ is convex. Then, by Lemma 2.1 in (Del Barrio and Loubes, 2019) we know ϕ is unique up to a constant. By Proposition 7.17 in (Santambrogio, 2015), we know the first variation of the Wasserstein term is $\frac{\phi}{\tau}$. The first variation of the potential energy term and the self-interaction term are U and $\int W(\cdot, y)\mathrm{d}\rho_\tau(y) + \int W(y, \cdot)\mathrm{d}\rho_\tau(y)$ respectively (see the first bullet point of Remark 7.13 in (Santambrogio, 2015)). Next, we will show the first variation of the internal energy functional is $h'(\rho_\tau)$.

Let $\tilde{\rho} = c = |\mathcal{X}|^{-1}$ be the constant positive density on $\mathcal{X}$, and let $\rho_\varepsilon = (1-\varepsilon)\rho_\tau + \varepsilon\tilde{\rho}$. Similar to Lemma 8.6 in (Santambrogio, 2015), we have $C_U\varepsilon \geqslant \varepsilon(\rho_\tau - c)h'(\rho_\varepsilon)$ where C_U is a positive value depending on U. Let

$$A = \{x \in \mathcal{X} : \rho_\tau(x) > 0\}, \quad B = \{x \in \mathcal{X} : \rho_\tau(x) = 0\}.$$

Then, we have

$$\begin{aligned} C_U \geqslant \int_{\mathcal{X}} (\rho_\tau - c)h'(\rho_\varepsilon) &= \int_A (\rho_\tau - c)h'\big((1-\varepsilon)\rho_\tau + \varepsilon c\big) - c|B|h'(\varepsilon c) \\ &\geqslant \int_A (\rho_\tau - c)h'(c) - c|B|h'(\varepsilon c). \end{aligned}$$

Here, we get the last inequality since h is convex and h' in nondecreasing. This implies $\rho_\tau \in L^1(\mathcal{X})$. Taking $\tau \to 0$ and applying Fatou's lemma, we have $(\rho_\tau - c)h'(\rho_\tau) \in L^1(\mathcal{X})$. Since $h(\rho_\tau) \in L^1(\mathcal{X})$ which implies $\rho_\tau h'(\rho_\tau) \in L^1(\mathcal{X})$, we know $h'(\rho_\tau) \in L^1(\mathcal{X})$. Therefore, for any $\bar{\rho} \in L^\infty(\mathcal{X})$, we have

$$\left|\frac{\partial}{\partial\varepsilon}h\big((1-\varepsilon)\rho_\tau + \varepsilon\bar{\rho}\big)\right| = \left|h'\big((1-\varepsilon)\rho_\tau + \varepsilon\bar{\rho}\big)\cdot(\bar{\rho} - \rho_\tau)\right| \leqslant \big(\|\bar{\rho}\|_\infty + \rho_\tau\big)\max\Big\{|h'(\rho_\tau)|, \|\bar{\rho}\|_\infty\Big\} \in L^1(\mathcal{X}).$$

Now, by Theorem 2.27 in (Folland, 1999) we know

$$\frac{\partial}{\partial\varepsilon}\int_{\mathcal{X}} h\big((1-\varepsilon)\rho_\tau + \varepsilon\bar{\rho}\big)\bigg|_{\varepsilon=0} = \int_{\mathcal{X}} \frac{\partial}{\partial\varepsilon}h\big((1-\varepsilon)\rho_\tau + \varepsilon\bar{\rho}\big)\bigg|_{\varepsilon=0} = \int_{\mathcal{X}} h'(\rho_\tau)\,\mathrm{d}(\bar{\rho} - \rho_\tau).$$

This indicates that $h'(\rho_\tau)$ is a subdifferential.

By Proposition 7.20 in Santambrogio (2015), we know

$$U + h'(\rho_\tau) + \frac{\phi}{\tau} + \int_{\mathcal{X}} W(\cdot, y)\,\mathrm{d}\rho_\tau(y) + \int_{\mathcal{X}} W(y, \cdot)\,\mathrm{d}\rho_\tau(y)$$

is a constant $[\rho_\tau]$-a.e.. Taking the gradient yields the result. ■

C.3 Proof of Lemma 12

Lemma 12. Let $\rho, \rho' \in \mathscr{P}_2^r(\mathcal{X})$. Under Assumption B, for any $j \in [m]$ we have

$$\left\| \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho_{-j} - \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho'_{-j} \right\|^2 \leqslant L^2 \, \mathrm{W}_2^2(\rho_{-j}, \rho'_{-j}).$$

Proof Just notice that

$$\begin{aligned}
&\left\| \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho_{-j} - \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho'_{-j} \right\|^2 \\
&= \left\| \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho_{-j} - \nabla \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\big(T_{\rho_{-j}}^{\rho'_{-j}}\big)_{\#}\rho_{-j} \right\|^2 \\
&= \left\| \int_{\mathcal{X}_{-j}} \nabla_j V(x_j, x_{-j}) - \nabla_j V\big(x_j, T_{\rho_{-j}}^{\rho'_{-j}}(x_{-j})\big) \, \mathrm{d}\rho_{-j} \right\|^2 \\
&\overset{\text{(i)}}{\leqslant} \int_{\mathcal{X}_{-j}} \left\| \nabla_j V(x_j, x_{-j}) - \nabla_j V\big(x_j, T_{\rho_{-j}}^{\rho'_{-j}}(x_{-j})\big) \right\|^2 \mathrm{d}\rho_{-j} \\
&\overset{\text{(ii)}}{\leqslant} L^2 \int_{\mathcal{X}_{-j}} \left\| T_{\rho_{-j}}^{\rho'_{-j}}(x_{-j}) - x_{-j} \right\|^2 \mathrm{d}\rho_{-j} \\
&= L^2 \, \mathrm{W}_2^2(\rho_{-j}, \rho'_{-j}).
\end{aligned}$$

Here, (i) is by Cauchy–Schwarz inequality; (ii) is by the L-smoothness of V. ∎

C.4 Tensorization of Wasserstein Distance

Lemma 13 (tensorization). Assume $\rho = \bigotimes_{j=1}^m \rho_j$ and $\rho' = \bigotimes_{j=1}^m \rho'_j$, where $\rho_j, \rho_j \in \mathscr{P}_2^r(\mathcal{X}_j)$ for all $j \in [m]$. Then

$$\mathrm{W}_2^2(\rho, \rho') = \sum_{j=1}^m \mathrm{W}_2^2(\rho_j, \rho'_j) \quad \text{and} \quad T_\mu^\nu(x) = (T_{\rho_1}^{\rho'_1}(x_1), \ldots, T_{\rho_m}^{\rho'_m}(x_m)).$$

Proof By definition, we have

$$\begin{aligned}
\mathrm{W}_2^2(\rho, \rho') &= \min \left\{ \int_{\mathcal{X}^2} \|x - y\|^2 \, \mathrm{d}\pi : \pi \in \Gamma(\rho, \rho') \right\} \\
&= \min \left\{ \sum_{j=1}^m \int_{\mathcal{X}_j^2} \|x_j - y_j\|^2 \, \mathrm{d}\pi(x, y) : \pi \in \Gamma(\rho, \rho') \right\} \\
&\geqslant \sum_{j=1}^m \min \left\{ \int_{\mathcal{X}_j^2} \|x_j - y_j\|^2 \, \mathrm{d}\pi(x, y) : \pi \in \Gamma(\rho, \rho') \right\} \\
&= \sum_{j=1}^m \min \left\{ \int_{\mathcal{X}_j^2} \|x_j - y_j\|^2 \, \mathrm{d}\pi_j(x_j, y_j) : \pi_j \in \Gamma(\rho_j, \rho'_j) \right\} = \sum_{j=1}^m \mathrm{W}_2^2(\rho_j, \rho'_j).
\end{aligned}$$

The inequality holds while taking $\pi = (\mathrm{Id}, T_\rho^{\rho'})_{\#}\rho$. This implies the desired result. ∎

C.5 Proof of Proposition 4

For any $\tilde{\rho}$, let $\tilde{T}$ be the optimal map from ρ^k to $\tilde{\rho}$. Note that

$$\begin{aligned}
&\mathcal{V}(\rho^{k+1}) + \mathcal{H}(\rho^{k+1}) + \mathcal{W}(\rho^{k+1}) + \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho^k, \rho^{k+1}) \\
\leqslant\ &\mathcal{V}(T_\#^{k+1}\rho^k) + \mathcal{H}(T_\#^{k+1}\rho^k) + \mathcal{W}(T_\#^{k+1}\rho^k) + \frac{1}{2\tau}\int \|T^{k+1}(x) - x\|^2\,\mathrm{d}\rho^k \\
\leqslant\ &\mathcal{V}(\tilde{T}_\#\rho^k) + \mathcal{H}(\tilde{T}_\#\rho^k) + \mathcal{W}(\tilde{T}_\#\rho^k) + \frac{1}{2\tau}\int \|\tilde{T}(x) - x\|^2\,\mathrm{d}\rho^k \\
=\ &\mathcal{V}(\tilde{\rho}) + \mathcal{H}(\tilde{\rho}) + \mathcal{W}(\tilde{\rho}) + \frac{1}{2\tau}\,\mathrm{W}_2^2(\rho^k, \tilde{\rho}).
\end{aligned}$$

So, $\rho^{k+1} = T_\#^{k+1}\rho^k$ minimizes (10).

C.6 Lemmas for Controlling Numerical Error

Lemma 14. Let $\{x_k\}_{k=0}^\infty$ and $\{\xi_k\}_{k=1}^\infty$ be two non-negative sequences such that

$$x_k \leqslant Ax_{k-1} + B\xi_k, \qquad \forall\, k \in \mathbb{Z}_+$$

for some constants $0 < A < 1$ and $B > 0$.

(1) If $\xi_k \leqslant \kappa\xi^k$ for some constants $0 < \xi < 1$ and $\kappa > 0$, we have

$$x_k \leqslant A^k x_0 + \frac{B\kappa}{|A - \xi|} \cdot \max\{A, \xi\}^{k+1}.$$

(2) If $\xi_k \leqslant \xi k^{-\alpha}$ for some $\xi, \alpha \geqslant 0$, there is a constant $C(\alpha, A) > 0$ such that

$$x_k \leqslant A^k x_0 + \frac{C(\alpha, A)B\xi}{k^\alpha}.$$

Proof It is easy to see that

$$x_k \leqslant A^k x_0 + B\sum_{l=1}^k A^{k-l}\xi_l, \qquad \forall\, k \in \mathbb{Z}_+.$$

(1) When $\xi_k \leqslant \kappa\xi^k$, we have

$$x_k \leqslant A^k x_0 + B\sum_{l=1}^k A^{k-l}\kappa\xi^l = A^k x_0 + BA^k\kappa\sum_{l=1}^K \left(\frac{\xi}{A}\right)^l.$$

Note that

$$\sum_{l=1}^k \left(\frac{\xi}{A}\right)^l \leqslant \begin{cases} \frac{A}{A-\xi} & \xi < A \\ \\ \frac{\xi^{k+1}}{A^k(\xi - A)} & \xi > A \end{cases}.$$

So, we have

$$BA^k\kappa\sum_{l=1}^{k}\left(\frac{\xi}{A}\right)^l \leqslant \begin{cases} \frac{AB\kappa}{A-\xi}A^k & \xi < A \\ \\ \frac{B\xi\kappa}{\xi-A}\xi^k & \xi > A \end{cases} \leqslant \frac{B\kappa}{|A-\xi|}\max\{A,\xi\}^{k+1}.$$

Therefore, we have

$$x_k \leqslant A^k x_0 + \frac{B\kappa}{|A-\xi|}\max\{A,\xi\}^{k+1} \leqslant \left(\frac{x_0}{A} + \frac{B\kappa}{|A-\xi|}\right)\max\{A,\xi\}^{k+1}.$$

(2) When $\xi_k \leqslant \xi k^{-\alpha}$, we have

$$x_k \leqslant A^k x_0 + B\sum_{l=1}^{k} A^{k-l}\frac{\xi}{l^\alpha} = A^k x_0 + BA^k\xi\sum_{l=1}^{k}\frac{A^{-l}}{l^\alpha}.$$

Similar to the proof in (Proof of Theorem 4, Yao et al., 2023), there is a constant $C = C(\alpha, A) > 0$ such that

$$\sum_{l=1}^{k}\frac{A^{-l}}{l^\alpha} \leqslant \frac{CA^{-k}}{k^\alpha}.$$

So, we have

$$x_k \leqslant A^k x_0 + BA^k\xi\frac{CA^{-k}}{k^\alpha} = A^k x_0 + \frac{C(\alpha, A)B\xi}{k^\alpha}.$$

■

Lemma 15. Let $\{x_k\}_{k=0}^{\infty}$ and $\{\xi_k\}_{k=1}^{\infty}$ be two non-negative sequences and $\{\xi_k\}$ are non-increasing. Assume that there are constants $A, B, C > 0$ such that

$$A{x_k}^2 + (1 - B\xi_k)x_k \leqslant x_{k-1} + C\xi_k^2.$$

If $\xi_k \leqslant \frac{\xi}{k^\alpha}$ for some $\alpha > 0$, then there exists a constant $K = K(\alpha, \xi, A, B, C) > 0$, such that

$$x_k \leqslant \frac{K}{k^{\min\{1,\alpha\}}}. \tag{28}$$

Proof Without loss of generality, we assume $\xi_k \leqslant \frac{1}{k^\alpha}$. Otherwise, we have

$$A{x_k}^2 + \big(1 - \frac{B\xi}{k^\alpha}\big)x_k \leqslant x_{k-1} + \frac{C\xi^2}{k^{2\alpha}},$$

and then take $B' = B\xi, C' = C\xi^2$, and $\xi_k' = 1/k^\alpha$.

We can take K large enough such that inequality (28) is true for all $k \geqslant k_0$, where $k_0 \geqslant 1$ is a constant such that $Bk_0^{-\alpha} < 1$. We will use induction to prove (28). Assume the inequality is true for $k - 1 \geqslant k_0$. Now let us prove (28) holds for k. Then, we have

$$x_k \leqslant \frac{-(1 - B\xi_k) + \sqrt{(1 - B\xi_k)^2 + 4A(x_{k-1} + C\xi_k^2)}}{2A}.$$

Since $1 - B\xi_k \geqslant 0$, we have

$$\begin{aligned} x_k \leqslant \frac{K}{k^{\min\{1,\alpha\}}} &\Longleftarrow \frac{-(1-B\xi_k) + \sqrt{(1-B\xi_k)^2 + 4A(x_{k-1} + C\xi_k^2)}}{2A} \leqslant \frac{K}{k^{\min\{1,\alpha\}}} \\ &\Longleftrightarrow (1-B\xi_k)^2 + 4A(x_{k-1} + C\xi_k^2) \leqslant \Big(\frac{2AK}{k^{\min\{1,\alpha\}}} + (1-B\xi_k)\Big)^2 \\ &\Longleftrightarrow x_{k-1} \leqslant \frac{AK^2}{k^{2\min\{1,\alpha\}}} + \frac{K(1-Bk^{-\alpha})}{k^{\min\{1,\alpha\}}} - \frac{C}{k^{2\alpha}} \\ &\Longleftrightarrow \frac{K}{(k-1)^{\min\{1,\alpha\}}} \leqslant \frac{AK^2}{k^{2\min\{1,\alpha\}}} + \frac{K(1-Bk^{-\alpha})}{k^{\min\{1,\alpha\}}} - \frac{C}{k^{2\alpha}} \\ &\Longleftrightarrow \frac{K}{(k-1)^{\min\{1,\alpha\}}} - \frac{K}{k^{\min\{1,\alpha\}}} \leqslant \frac{AK^2}{k^{2\min\{1,\alpha\}}} - \frac{BK}{k^{\alpha+\min\{1,\alpha\}}} - \frac{C}{k^{2\alpha}}. \end{aligned}$$

Case 1. When $\alpha > 1$, we only need to prove

$$\frac{K}{k-1} - \frac{K}{k} \leqslant \frac{AK^2}{k^2} - \frac{BK}{k^{1+\alpha}} - \frac{C}{k^{2\alpha}} \Longleftarrow \frac{K}{k(k-1)} \leqslant \frac{AK^2 - BK - C}{k^2},$$

which only requires $AK^2 - (B+2)K - C \geqslant 0$ since $k \geqslant k_0 + 1 \geqslant 2$.
Case 2. When $\alpha \leqslant 1$, we only need to show

$$\frac{K}{(k-1)^\alpha} - \frac{K}{k^\alpha} \leqslant \frac{AK^2 - BK - C}{k^{2\alpha}}.$$

Note that

$$k^\alpha\Big[\frac{1}{(k-1)^\alpha} - \frac{1}{k^\alpha}\Big] = \Big(1 + \frac{1}{k-1}\Big)^\alpha - 1 \leqslant 2^{(\alpha-1)_+}\frac{\alpha}{k-1} \leqslant \frac{2^\alpha\alpha}{k} \qquad \forall\, k \geqslant 2,$$

So, we only need to prove

$$\frac{2^\alpha\alpha K}{k} \leqslant \frac{AK^2 - BK - C}{k^\alpha}.$$

Since $\alpha < 1$, we just need $AK^2 - (B + 2^\alpha\alpha)K - C \geqslant 0$. By induction, we know such K exists. ■

C.7 Numerical Methods in FA Approach

By changing variables, we can write the interaction and self-interaction functionals as

$$\mathcal{V}(T_\#\rho^k) = \int V(x)\,\mathrm{d}[T_\#\rho^k](x) = \int V(T(x))\,\mathrm{d}\rho^k(x),$$

and

$$\mathcal{W}(T_\#\rho^k) = \iint W(x,x')\,\mathrm{d}[T_\#\rho^k](x)\mathrm{d}[T_\#\rho^k](x') = \iint W(T(x),T(x'))\,\mathrm{d}\rho^k(x)\mathrm{d}\rho^k(x').$$

These expressions suggest the following sample approximations

$$\mathcal{V}(T_{\#}\rho^k) \approx \frac{1}{B}\sum_{b=1}^{B} V(T(X_b^k)) \quad \text{and} \quad \mathcal{W}(T_{\#}\rho^k) \approx \frac{1}{B^2}\sum_{b,b'=1}^{B} W\big(T(X_b^k), T(X_{b'}^k)\big),$$

where $\{X_b^k : b \in [B]\}$ are B samples from ρ^k. Regarding the internal energy term $\mathcal{H}(\rho^k)$, we will consider two leading cases $\mathcal{H}(\rho^k) = \int \rho_k \log \rho_k$ and $\mathcal{H}(\rho^k) = \int (\rho^k)^n$ with some positive integer $n \geqslant 2$.

For $\mathcal{H}(\rho^k) = \int \rho^k \log \rho^k$, we can apply the argument in (Mokrov et al., 2021) by noting that

$$\mathcal{H}(T_{\#}\rho^k) = \mathcal{H}(\rho^k) - \int \log |\det \nabla T(x)| \, \mathrm{d}\rho^k,$$

which suggests approximating the negative self-entropy functional by

$$\mathcal{H}(T_{\#}\rho^k) \approx \mathcal{H}(\rho^k) - \frac{1}{B}\sum_{b=1}^{B} \log |\det \nabla T(X_b^k)|$$

Since the first term does not depend on T, the above discussions yield the approximation in Example 5

For $\mathcal{H}(\rho^k) = \int (\rho^k)^n$, an additional kernel density estimation (KDE) step is required. Since

$$\mathcal{H}(T_{\#}\rho^k) = \int \big[\rho^k(x) \cdot |\det \nabla T(x)|^{-1}\big]^{n-1} \, \mathrm{d}\rho^k,$$

which suggests that we consider the approximation

$$\mathcal{H}(T_{\#}\rho^k) \approx \frac{1}{B}\sum_{b=1}^{B} \big[\widehat{\rho}_{\mathrm{kde}}^k(X_b^k) \cdot |\det \nabla T(X_b^k)|^{-1}\big]^{n-1},$$

where $\widehat{\rho}_{\mathrm{kde}}^k$ is the kernel density estimation of ρ^k by $\{X_b^k : b \in [B]\}$. This leads to the approximation in Example 6.

Appendix D. Results for Multi-species Systems

D.1 Proof of Proposition 25

Since $\rho^* = (\rho_1^*, \ldots, \rho_m^*)$ is the minimum of (1), by Lemma 9, we have

$$\begin{aligned} C_j(\rho_{-j}^*) &= \int_{\mathcal{X}_{-j}} V(x_j, x_{-j}) \, \mathrm{d}\rho_{-j}^* + h_j'(\rho_j^*) + \int_{\mathcal{X}_j} W_j(\cdot, y) \, \mathrm{d}\rho_j^*(y) + \int_{\mathcal{X}_j} W_j(y, \cdot) \, \mathrm{d}\rho_j^*(y) \\ &= \Big(V_j - \sum_{i \neq j} K_{ij} * \rho_i^*\Big) + h_j'(\rho_j^*) - K_{jj} * \rho_j^* \end{aligned}$$

is a constant $[\rho_j^*]$-a.e. on $\mathcal{X}_j$. If $\rho(t) = \rho^*$ at time $t = 0$, we have

$$\rho_j(t) \Big[\nabla V_j(t) - \sum_{i=1}^{m} (\nabla K_{ij}) * \rho_i + \nabla h_j'(\rho_j) \Big] = 0. \tag{29}$$

This implies that ρ^* is a stationary measure.

D.2 Example in Section 5.2

Recall that $h_j(\rho_j) = \rho_j \log \rho_j$,

$$V(x_1, x_2, x_3) = \sum_{i=1}^{n} \frac{r_i}{2} \|x_i - m_i\|^2 - \sum_{1 \leqslant i < j \leqslant 3} \frac{Q_i Q_j}{2} \arctan \|x_i - x_j\|^2$$

and

$$W_j(x_j, x_j') = \frac{Q_j^2}{2} \arctan \|x_j - x_j'\|^2, \quad 1 \leqslant j \leqslant 3.$$

Let us now check all assumptions in Section 3.
Assumption A. It is clear that ρ_j^2 satisfies the Assumption A.

Assumption B. Let $R(y) = y/(1 + \|y\|^2)$ be a function on $\mathbb{R}^2$. For any $x = (x_1, x_2, x_3)$ and $x' = (x_1', x_2', x_3')$, we have

$$\begin{aligned}
\big\|\nabla_1 V(x) - \nabla_1 V(x')\big\| \lesssim{}& \big\|r_1(x_1 - x_1')\big\| + |Q_1 Q_2| \cdot \Big\| \frac{x_1 - x_2}{1 + \|x_1 - x_2\|^2} - \frac{x_1' - x_2'}{1 + \|x_1' - x_2'\|^2} \Big\| \\
&+ |Q_1 Q_3| \cdot \Big\| \frac{x_1 - x_3}{1 + \|x_1 - x_3\|^2} - \frac{x_1' - x_3'}{1 + \|x_1' - x_3'\|^2} \Big\| \\
={}& |r_1| \cdot \|x_1 - x_1'\| + |Q_1 Q_2| \cdot \big\|R(x_1 - x_2) - R(x_1' - x_2')\big\| \\
&+ |Q_1 Q_3| \cdot \big\|R(x_1 - x_3) - R(x_1' - x_3')\big\|.
\end{aligned}$$

Notice that the Jacobian matrix of R at $y = (y_1, y_2)$ is

$$JR(y) = \frac{1}{(1 + y_1^2 + y_2^2)^2} \begin{pmatrix} 1 + y_2^2 - y_1^2 & -2y_1 y_2 \\ -2y_1 y_2 & 1 + y_1^2 - y_2^2 \end{pmatrix}$$

with eigenvalues

$$\lambda_1 = \frac{1}{1 + y_1^2 + y_2^2}, \quad \text{and} \quad \lambda_2 = \frac{1 - y_1^2 - y_2^2}{(1 + y_1^2 + y_2^2)^2}.$$

Clearly, we have $|\lambda_1| \leqslant 1$ and $|\lambda_2| \leqslant 1$. This implies $\|R(y) - R(y')\| \leqslant \|y - y'\|$. Therefore, we have

$$\begin{aligned}
\big\|\nabla_1 V(x) - \nabla_1 V(x')\big\| &\lesssim \|x_1 - x_1'\| + \|(x_1 - x_2) - (x_1' - x_2')\| + \|(x_1 - x_3) - (x_1' - x_3')\| \\
&\lesssim \|x - x'\|.
\end{aligned}$$

This indicates that

$$\big\|\nabla V(x) - \nabla V(x')\big\| \lesssim \sum_{i=1}^{3} \big\|\nabla_i V(x) - \nabla_i V(x')\big\| \lesssim \|x - x'\|.$$

So, ∇V is Lipschitz.

Assumption C. To verify this, we first prove that when $\alpha_j = \beta_j = Q_j^2$,

$$\widetilde{W}_j(x_j, x_j') = \frac{\alpha_j}{2}\|x_j\|^2 + W_j(x_j, x_j') + \frac{\beta_j}{2}\|x_j'\|^2 = \frac{\alpha_j}{2}\|x_j\|^2 + \frac{Q_j^2}{2}\arctan\|x_j - x_j'\|^2 + \frac{\beta_j}{2}\|x_j'\|^2$$

is a convex function on $\mathbb{R}^2 \times \mathbb{R}^2$. To show this, notice that

$$\frac{\partial W_j}{\partial x_j} = Q_j^2 R(x_j - x_j') \quad \text{and} \quad \frac{\partial W_j}{\partial x_j'} = Q_j^2 R(x_j' - x_j).$$

Therefore, the Hessian of W_j is

$$\nabla^2 W_j = Q_j^2 \begin{pmatrix} JR(x_j - x_j') & -JR(x_j - x_j') \\ -JR(x_j - x_j') & JR(x_j - x_j') \end{pmatrix}.$$

Let

$$\eta_{j1} = \frac{1}{1 + (x_{j1} - x_{j1}')^2 + (x_{j2} - x_{j2}')^2} \quad \text{and} \quad \eta_{j2} = \frac{1 - (x_{j1} - x_{j1}')^2 - (x_{j2} - x_{j2}')^2}{(1 + (x_{j1} - x_{j1}')^2 + (x_{j2} - x_{j2}')^2)^2}$$

be the eigenvalues of $JR(x_j - x_j')$. Assume $P \in O(2)$ be the orthogonal transition matrix such that

$$JR(x_j - x_j') = P \begin{pmatrix} \eta_{j1} & 0 \\ 0 & \eta_{j2}. \end{pmatrix} P^{\mathrm{T}}.$$

Therefore, we have

$$\begin{pmatrix} P^{\mathrm{T}} & 0 \\ 0 & P^{\mathrm{T}} \end{pmatrix} \nabla^2 \widetilde{W}_j \begin{pmatrix} P & 0 \\ 0 & P \end{pmatrix} = \begin{pmatrix} \alpha_j + Q_j^2\eta_{j1} & 0 & -\eta_{j1} & 0 \\ 0 & \alpha_j + Q_j^2\eta_{j2} & 0 & -\eta_{j2} \\ -\eta_{j1} & 0 & \beta_j + Q_j^2\eta_{j1} & 0 \\ 0 & -\eta_{j2} & 0 & \beta_j + Q_j^2\eta_{j2} \end{pmatrix},$$

which is p.s.d. when $\alpha_j = \beta_j \geqslant -Q_j^2 \min\{0, \eta_{j1}, \eta_{j2}\}$. We can then take $\alpha_j = \beta_j = Q_j^2$ since $\eta_{j1}, \eta_{j2} > -1$.

Next, we will show that

$$\begin{aligned} \widetilde{V}(x) &= V(x) - \sum_{i=1}^{3} Q_i^2 \|x_i\|^2 \\ &= \sum_{i=1}^{n} \left(\frac{r_i}{2}\|x_i - m_i\|^2 - Q_i^2\|x_i\|^2 \right) - \sum_{1 \leqslant i < j \leqslant 3} \frac{Q_i Q_j}{2} \arctan\|x_i - x_j\|^2 \end{aligned}$$

is λ-convex for $\lambda = \min_{1 \leqslant i \leqslant 3} \left(r_i - 4Q_i^2 - |Q_i| \sum_{j \neq i} |Q_j| \right)$. Let $H_{ij} = JR(x_i - x_j)$ for simplicity. It is easy to see that

$$\frac{\partial^2 \widetilde{V}}{\partial x_i^2} = (r_i - 2Q_i^2) I_2 - Q_i \sum_{l \neq i} Q_l H_{il} \quad \text{and} \quad \frac{\partial \widetilde{V}}{\partial x_i \partial x_j} = Q_i Q_j H_{ij}, \quad \forall\, 1 \leqslant i \neq j \leqslant 3.$$

Therefore, we have

$$
\begin{aligned}
&\begin{pmatrix} y_1^{\mathrm{T}} & y_2^{\mathrm{T}} & y_3^{\mathrm{T}} \end{pmatrix} \nabla^2 \widetilde{V}(x) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \\
&= \sum_{i=1}^3 (r_i - 2Q_i^2)\|y_i\|^2 - \sum_{i=1}^3 Q_i \sum_{j\neq i} Q_j y_i^{\mathrm{T}} H_{ij} y_i + 2 \sum_{1\leqslant i<j\leqslant 3} Q_i Q_j y_i^{\mathrm{T}} H_{ij} y_j \\
&\geqslant \sum_{i=1}^3 (r_i - 2Q_i^2)\|y_i\|^2 - \sum_{i=1}^3 \sum_{j\neq i} |Q_i Q_j| \|\!|H_{ij}|\!\|_{\mathrm{op}} \|y_i\|^2 - 2 \sum_{1\leqslant i<j\leqslant 3} |Q_i Q_j| \|\!|H_{ij}|\!\|_{\mathrm{op}} \|y_i\| \|y_j\| \\
&\overset{\text{(i)}}{\geqslant} \sum_{i=1}^3 \Big(r_i - 2Q_i^2 - |Q_i| \sum_{j\neq i} |Q_j| \Big) \|y_i\|^2 - 2 \sum_{1\leqslant i<j\leqslant 3} |Q_i Q_j| \|y_i\| \|y_j\| \\
&\overset{\text{(ii)}}{\geqslant} \sum_{i=1}^3 \Big(r_i - 4Q_i^2 - |Q_i| \sum_{j\neq i} |Q_j| \Big) \|y_i\|^2 \\
&\geqslant \min_{1\leqslant i\leqslant 3} \Big(r_i - 4Q_i^2 - |Q_i| \sum_{j\neq i} |Q_j| \Big) \|y\|^2
\end{aligned}
$$

Here, (i) is because $\|\!|H_{ij}|\!\|_{\mathrm{op}} \leqslant 1$; (ii) is due to AM-GM inequality. This implies $\widetilde{V}$ is $\min_{1\leqslant i\leqslant 3} \Big(r_i - 4Q_i^2 - |Q_i| \sum_{j\neq i} |Q_j| \Big)$-convex.

Appendix E. Connections between Different Lipschitz Conditions

First, let us show $0 \leqslant L/L_c \leqslant \sqrt{m-1}$. The lower bound is because L and L_c are non-negative. For the upper bound, notice that $L = \max_{j\in[m]} \sup_x \|\!|\nabla_j \nabla_{-j} V(x)|\!\|_{\mathrm{op}}$ and $L_c = \sup_x \left\|\!\left|\nabla_j^2 V(x)\right|\!\right\|_{\mathrm{op}}$. Also, when $\nabla_j^2 V$ is invertible, we have

$$
0 \leq \begin{pmatrix} \nabla_{-j}^2 V & \nabla_j \nabla_{-j} V \\ \nabla_{-j} \nabla_j V & \nabla_j^2 V \end{pmatrix} \sim \begin{pmatrix} \nabla_j^2 V & 0 \\ 0 & \nabla_{-j}^2 V - \nabla_j \nabla_{-j} V (\nabla_j^2 V)^{-1} \nabla_{-j} \nabla_j V \end{pmatrix},
$$

where $A \sim B$ means A and B are congruent, i.e., there exists invertible matrix P such that $A = P^T B P$. It is known that congruent transformation does not change the semi-positive definiteness of a matrix. This implies $0 \leq \nabla_j^2 V - \nabla_j \nabla_{-j} V (\nabla_j^2 V)^{-1} \nabla_{-j} \nabla_j V$, i.e.,

$$
v^T \nabla_{-j}^2 V v \geqslant v^T \nabla_j \nabla_{-j} V (\nabla_j^2 V)^{-1} \nabla_{-j} \nabla_j V v, \quad \forall\, v \in \mathbb{R}^{d-d_j}.
$$

Therefore, we have

$$
\begin{aligned}
\|\!|\nabla_{-j}^2 V|\!\|_{\mathrm{op}} &= \sup_{\|v\|=1} v^T \nabla_{-j}^2 V v \\
&\geqslant \sup_{\|v\|=1} v^T \nabla_j \nabla_{-j} V (\nabla_j^2 V)^{-1} \nabla_{-j} \nabla_j V v \\
&\geqslant \sup_{\|v\|=1} \|\!|\nabla_j^2 V|\!\|_{\mathrm{op}}^{-1} \|\nabla_{-j} \nabla_j V v\|^2
\end{aligned}
$$

$$= \left\|\left\|\nabla_j^2 V\right\|\right\|_{\mathrm{op}}^{-1} \left\|\left\|\nabla_{-j}\nabla_j V\right\|\right\|_{\mathrm{op}}^2.$$

This implies

$$\begin{aligned}
L^2 = \sup_x \left\|\left\|\nabla_{-j}\nabla_j V(x)\right\|\right\|_{\mathrm{op}}^2 &\leqslant \sup_x \left\|\left\|\nabla_{-j}^2 V(x)\right\|\right\|_{\mathrm{op}} \left\|\left\|\nabla_j^2 V(x)\right\|\right\|_{\mathrm{op}} \\
&\leqslant \sup_x \left\|\left\|\nabla_{-j}^2 V(x)\right\|\right\|_{\mathrm{op}} \sup_x \left\|\left\|\nabla_j^2 V(x)\right\|\right\|_{\mathrm{op}} \\
&\overset{\text{(i)}}{\leqslant} (m-1) \max_{i\neq j} \sup_x \left\|\left\|\nabla_i^2 V(x)\right\|\right\|_{\mathrm{op}} \sup_x \left\|\left\|\nabla_j^2 V(x)\right\|\right\|_{\mathrm{op}} \\
&\leqslant (m-1)\Big(\max_{j\in[m]} \sup_x \left\|\left\|\nabla_j^2 V(x)\right\|\right\|_{\mathrm{op}}\Big)^2 \\
&= (m-1)L_c^2.
\end{aligned}$$

Here, (i) is due to Section 3.2 in (Wright, 2015). Thus, we have $L/L_c \leqslant \sqrt{m-1}$. To see the bounds are tight, consider the examples $V(x_1, x_2) = x_1^2 + x_2^2$ and $V(x_1, \cdots, x_n) = (x_m + \cdots, x_m)^2$. In the first example, we have $L = 0$ and $L_c = 2$, indicating that $L/L_c = 0$. In the second example, we have $L = 2\sqrt{m-1}$ and $L_c = 2$, indicating that $L/L_c = \sqrt{m-1}$.

Next, we will prove that $0 \leqslant L/L_r \leqslant \sqrt{1-1/m}$. The lower bound is due to $L, L_r \geqslant 0$ and is tight due to the same reason as of L/L_c. To see the upper bound, note that $L_r = \max_{j\in[m]} \left\|\left\|\nabla_j \nabla V\right\|\right\|_{\mathrm{op}}$. Therefore, we have

$$\begin{aligned}
L_r^2 &= \max_{j\in[m]} \sup_{\|v_1\|^2+\|v_2\|^2=1} \|\nabla_{-j}\nabla_j V v_1 + \nabla_j^2 V v_2\|^2 \\
&\overset{\text{(i)}}{=} \max_{j\in[m]} \sup_{\|v_1\|^2+\|v_2\|^2=1} \Big(\|\nabla_{-j}\nabla_j V v_1\| + \|\nabla_j^2 V v_2\|\Big)^2 \\
&\overset{\text{(ii)}}{=} \max_{j\in[m]} \sup_{\|v_1\|^2+\|v_2\|^2=1} \Big(\left\|\left\|\nabla_{-j}\nabla_j V\right\|\right\|_{\mathrm{op}}\|v_1\| + \left\|\left\|\nabla_j^2 V\right\|\right\|_{\mathrm{op}}\|v_2\|\Big)^2 \\
&\overset{\text{(iii)}}{\geqslant} \sup_{\|v_1\|^2+\|v_2\|^2=1} \Big(L\|v_1\| + \frac{L}{\sqrt{m-1}}\|v_2\|\Big)^2 \\
&= \frac{mL^2}{m-1}.
\end{aligned}$$

Here, (i) is because we can rotate v_2 such that $\nabla_{-j}\nabla_j V v_1$ and $\nabla_j^2 V v_2$ have the same direction; (ii) is by the definition of operator norm; (iii) is by $L_c \geqslant L/\sqrt{m-1}$. Thus, we have $L/L_r \leqslant \sqrt{1-1/m}$. The tightness of this bound can been seen by considering $V(x) = (x_1 + \cdots + x_n)^2$.

Finally, let us show $0 \leqslant L/L_{\mathrm{g}} \leqslant 1$. This inequality is obviously true. The lower bound is tight by considering $V(x) = x_1^2 + \cdots + x_m^2$. For the upper bound, consider the example $V(x) = (x_1 + x_2)^2 + x_3^2 + \cdots + x_m^2$. In this example, we have $L = L_{\mathrm{g}} = 2$. Therefore, $L/L_{\mathrm{g}} \leqslant 1$ is the tightest bound.

References

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

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/amos17b.html`.

Manuel Arnese and Daniel Lacker. Convergence of coordinate ascent variational inference for log-concave measures via optimal transport. *arXiv preprint arXiv:2404.08792*, 2024.

Donald G Aronson. The porous medium equation. *Nonlinear Diffusion Problems: Lectures given at the 2nd 1985 Session of the Centro Internazionale Matermatico Estivo (CIME) held at Montecatini Terme, Italy June 10–June 18, 1985*, pages 1–46, 2006.

Marin Ballu and Quentin Berthet. Mirror sinkhorn: Fast online optimization on transport polytopes. *arXiv preprint arXiv:2211.10420*, 2022.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Adrien Blanchet, Pascal Mossay, and Filippo Santambrogio. Existence and uniqueness of equilibrium for a spatial model of social interactions. *International Economic Review*, 57 (1):31–60, 2016.

François Bolley. Separability and completeness for the wasserstein distance. In *Séminaire de probabilités XLI*, pages 371–377. Springer, 2008.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Roberto Carlos Cabrales, Juan Vicente Gutiérrez-Santacreu, and José Rafael Rodríguez-Galván. Numerical solution for an aggregation equation with degenerate diffusion. *Applied Mathematics and Computation*, 377:125145, 2020.

Jérôme Carrayrou, Robert Mosé, and Philippe Behra. New efficient algorithm for solving thermodynamic chemistry. *AIChE Journal*, 48(4):894–904, 2002.

José A Carrillo, Yanghong Huang, and Markus Schmidtchen. Zoology of a nonlocal cross-diffusion model for two species. *SIAM Journal on Applied Mathematics*, 78(2):1078–1104, 2018.

José A Carrillo, Katy Craig, and Yao Yao. Aggregation-diffusion equations: dynamics, asymptotics, and singular limits. *Active Particles, Volume 2: Advances in Theory, Models, and Applications*, pages 65–108, 2019.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

Esther S Daus, Markus Fellner, and Ansgar Jüngel. Random-batch method for multi-species stochastic interacting particle systems. *Journal of Computational Physics*, 463:111220, 2022.

Eustasio Del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

Marco Di Francesco and Simone Fagioli. Measure solutions for non-local interaction pdes with two species. *Nonlinearity*, 26(10):2777, 2013.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. 2017.

Joep HM Evers, Razvan C Fetecau, and Theodore Kolokolnikov. Equilibria for an aggregation model with two species. *SIAM Journal on Applied Dynamical Systems*, 16(4): 2287–2338, 2017.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Soumyadip Ghosh, Yingdong Lu, Tomasz Nowicki, and Edith Zhang. On representations of mean-field variational inference. *arXiv preprint arXiv:2210.11385*, 2022.

Wenxuan Guo, YoonHaeng Hur, Tengyuan Liang, and Chris Ryan. Online learning to transport via the minimal selection principle. In *Conference on Learning Theory*, pages 4085–4109. PMLR, 2022.

Timon Gutleb, José Carrillo, and Sheehan Olver. Computing equilibrium measures with power law kernels. *Mathematics of Computation*, 91(337):2247–2281, 2022.

Xiaoqin Hua and Nobuo Yamashita. Iteration complexity of a block coordinate gradient descent method for convex optimization. *SIAM Journal on Optimization*, 25(3):1298–1313, 2015.

Shi Jin, Lei Li, and Jian-Guo Liu. Random batch methods (rbm) for interacting particle systems. *Journal of Computational Physics*, 400:108877, 2020.

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.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*, 2022.

Hugo Lavenant and Giacomo Zanella. Convergence rate of random scan coordinate ascent variational inference under log-concavity. *arXiv preprint arXiv:2406.07292*, 2024.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

Dong Li and José L Rodrigo. Wellposedness and regularity of solutions of an aggregation equation. *Revista Matemática Iberoamericana*, 26(1):261–294, 2010.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *The Journal of Machine Learning Research*, 18(1):6741–6764, 2017.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

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

Kolade M Owolabi and Abdon Atangana. Computational study of multi-species fractional reaction-diffusion system with abc operator. *Chaos, Solitons & Fractals*, 128:280–289, 2019.

Juan Manuel Paz-Garcia, Björn Johannesson, Lisbeth M Ottosen, Alexandra B Ribeiro, and José Miguel Rodríguez-Maroto. Computing multi-species chemical equilibrium with an algorithm based on the reaction extents. *Computers & chemical engineering*, 58:135–143, 2013.

Zehra Pinar. The reaction–cross-diffusion models for tissue growth. *Mathematical Methods in the Applied Sciences*, 44(18):13805–13811, 2021.

R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

Ruoyu Sun and Yinyu Ye. Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version. *Mathematical Programming*, 185:487–520, 2021.

Nicolas Garcia Trillos and Daniel Sanz-Alonso. The bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.

Juan Luis Vázquez. *The porous medium equation: mathematical theory*. Oxford University Press on Demand, 2007.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015.

Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*, 2023.

Rentian Yao and Yun Yang. Mean field variational inference via wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*, 2022.

Rentian Yao, Linjun Huang, and Yun Yang. Minimizing convex functionals over space of probability measures via kl divergence gradient flow. *arXiv preprint arXiv:2311.00894*, 2023.

Nicola Zamponi and Ansgar Jüngel. Analysis of degenerate cross-diffusion population models with volume filling. In *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire*, volume 34, pages 1–29. Elsevier, 2017.