Journal of Machine Learning Research 26 (2025) 1-46 Submitted 5/24; Revised 7/25; Published 10/25

Mean-Field Variational Inference via Wasserstein Gradient Flow

Rentian Yao RENTIAN2@MATH.UBC.CA
Department of Mathematics
University of British Columbia
Vancouver, BC V6S-0M7, CA

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

Editor: Quentin Berthet

Abstract

Variational inference, such as the mean-field (MF) approximation, requires certain conjugacy structures for efficient computation. These can impose unnecessary restrictions on the viable prior distribution family and further constraints on the variational approximation family. In this work, we introduce a general computational framework to implement MF variational inference for Bayesian models, with or without latent variables, using the Wasserstein gradient flow (WGF), a modern mathematical technique for realizing a gradient flow over the space of probability measures. Theoretically, we analyze the algorithmic convergence of the proposed approaches, providing an explicit expression for the contraction factor. We also strengthen existing results on MF variational posterior concentration from a polynomial to an exponential contraction, by utilizing the fixed point equation of the time-discretized WGF. Computationally, we propose a new constraint-free function approximation method using neural networks to numerically realize our algorithm. This method is shown to be more precise and efficient than traditional particle approximation methods based on Langevin dynamics.

Keywords: Bayesian statistics, mean-field variational inference, optimal transport

1. Introduction

One of the core problems of modern Bayesian inference is to compute the posterior distribution, a joint probability measure over unknown quantities, such as model parameters and unobserved latent variables, obtained by combining data information with prior knowledge in a principled manner. Modern statistics often rely on complex models for which the posterior distribution is analytically intractable and requires approximate computation. As a common alternative strategy to conventional Markov chain Monte Carlo (MCMC) sampling approach for approximating the posterior, variational inference (VI, Bishop and Nasrabadi (2006)), or variational Bayes Fox and Roberts (2012), finds the closest member in a user specified class of analytically tractable distributions, referred to as the variational (distribution) family, to approximate the target posterior. Although MCMC is asymptotically exact, VI is usually orders of magnitude faster Blei et al. (2017); Salimans et al. (2015)

since it turns the sampling or integration into an optimization problem. VI has successfully demonstrated its power in a wide variety of applications, including clustering Blei and Jordan (2006); Corduneanu and Bishop (2001), semi-supervised learning Kingma et al. (2014), neural-network training Anderson and Peterson (1987); Opper and Winther (1997), and probabilistic modeling Blei et al. (2003); Jordan et al. (1999). Among various approximating schemes, the mean-field (MF) approximation, which originates from statistical mechanics and uses the approximating family consisting of all fully factorized density functions over (blocks of) the unknown quantities, is the most widely used and representative instance of VI that is conceptually simple yet practically powerful.

On the downside, VI still requires certain conditional conjugacy structure to facilitate efficient computation (c.f. Section 2.5), in the same spirit as the requirement of a closed-form E-step in the expectation-maximization (EM, Dempster et al. (1977)) algorithm, a famous iterative method for parameter estimation in statistical models involving unobserved latent variables. Such a requirement unfortunately may: 1. add restrictions to the viable prior distribution family, limiting the applicability of VI; 2. call for specifically designed tricks for the implementation, making the VI methodology less generic and user-friendly; 3. need impose further constraints on the variational family, leading to increased approximation error. For example, when implementing Bayesian Gaussian mixture models for clustering, although independent Gaussian priors of cluster centers meet the aforementioned conditional conjugacy property, it is sensible to instead employ a class of repulsive priors Xie and Xu (2020) to encourage the well-separatedness of cluster centers and reduce the potential redundancy of components. Unfortunately, the complicated dependence structure introduced by the repulsive prior destroys the conditional conjugacy, making the standard coordinate ascent variational inference (CAVI, Bishop and Nasrabadi (2006)) algorithm for implementing the MF approximation inapplicable (see Section 6.2 for further details). Another example is Bayesian logit model Jaakkola and Jordan (1997). Due to the lack of conditional conjugacy, Jaakkola and Jordan (2000) proposes to use a tangent transformation motivated by convex duality to make the variational approximation computationally tractable. For the mixed multinomial logit model, Braun and McAuliffe (2010) derives a variational procedure based on the multivariate delta method for moments, which again requires specialized treatments and lacks generality.

In this paper, we propose a new computational framework for MF variational inference based on Wasserstein gradient flow, that is, running a "gradient descent" over the Wasserstein space, the space of all probability distributions with finite second moments endowed with the 2-Wasserstein metric W_2 (Ambrosio et al., 2008). Comparing to existing approaches, our approach does not impose any extra restrictions on the MF variational family, and can be applied to Bayesian models without any structural constraint on the prior and data likelihood function.

1.1 Related work

There are a number of studies aiming at building generic VI procedures for dealing with non-conjugate models while maintaining the computational tractability. For example, Wang and Blei (2012) develops two generic methods, Laplace variational inference and delta method variational inference, for a class of non-conjugate models with certain constraints (more

precisely, partly-conjugate models), by enforcing the variational family for the model parameters in the MF approximation to be the (multivariate) location-scale Gaussian family. Automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017; Wingate and Weber, 2013) provides an automatic scheme that derives an iterative algorithm for implementing the variational inference based on automatic differentiation and stochastic gradient ascent; but the performance of ADVI heavily depends on the parametrization of the variational family, and little theory has been developed to analyze its algorithmic convergence. In a related thread, Ranganath et al. (2014) proposes black box variational inference (BBVI) based on stochastic optimization, which is shown to have exponential convergence up to the noise level of stochastic gradient. However, both ADVI and BBVI only apply to parametric variational families that are finite-dimensional, which may unnecessarily impose additional constraints on top of the MF approximation — the constituting components of the density product in the MF family should be restriction-free and may not be characterized by a finite number of parameters.

The notion of Wasserstein gradient flow is first introduced in the influential work of Jordan et al. (1998). The authors reveal an appealing connection between: 1. the dynamics of a gradient flux, or steepest descent, for minimizing the free energy with respect to the Wasserstein metric; 2. a special class of partial differential equations (PDE), called the Fokker-Planck equation Gardiner et al. (1985); Risken (1996), which describes the evolution of the probability density for the position of a particle whose motion is described by a corresponding Ito stochastic differential equation (SDE). Specifically, Jordan et al. (1998) constructs a discrete and iterative variational scheme, also called the Jordan-Kinderlehrer-Otto (JKO) scheme, which extends from the Euclidean gradient descent and whose solutions (weakly) converge to the solution of the Fokker-Planck equation with the gradient of a potential as the drift term. Later, Otto (2001) extends this connection to the porous medium equation, and points out the resemblance between the Wasserstein space and an "infinite-dimenisonal" Riemannian manifold. A comprehensive development of gradient flows in a general metric space, including the Wasserstein space as a representative application, is provided in the monograph Ambrosio et al. (2008). The deep connection between Wasserstein gradient flows and a rich class of PDE (SDE) builds a bridge between geometric analysis, optimal transport, control theory and partial differential equations; and also motivates a class of particle based methods Carrillo et al. (2022, 2019); Frogner and Poggio (2020) for numerically solving PDE (SDE). It is worth highlighting that the development of Wasserstein gradient flows heavily relies on recent techniques from modern optimal transport theory Brenier (1991); Monge (1781); Villani (2003, 2009).

Some recent works also apply gradient flow over the space of probability measures to faciliate the computation of Bayesian statistics. For example, Trillos and Sanz-Alonso (2020) considers sampling from the posterior distribution based on gradient flows in a different context, by treating the posterior distribution as the minimizer of functionals with certain forms; and they propose to use the gradient flow to guide the choice of proposals for MCMC methods. While we are preparing the manuscript, we learnt that a concurrent work Lambert et al. (2022) also study the application of Wasserstein gradient flow to the computation of variational inference. Unlike our work, Lambert et al. (2022) focuses on Gaussian variational inference, where the target posterior (without latent variables) is approximated by the closest member in the Gaussian (local-scale) distribution family. Since

the Gaussian distribution family is a parametric family, their gradient flow is defined on the Bures-Wasserstein space of Gaussian measures and is intrinsically finite-dimensional. Lambert et al. (2022) proves the exponential convergence of a time-discretized version of the evolutionary ODE on the mean vector and covariance matrix, under the assumption that the target posterior distribution is strictly log-concave. In contrast, the mean-field (MF) variational approximation considered in our work involves an infinite-dimensional family, and our Bayesian latent variable model framework accommodates latent variables that are of discrete types. Moreover, we also study the large-sample statistical properties of the MF approximation, utilizing the fixed-point equation of our proposed time-discretized Wasserstein gradient flow.

1.2 Contribution summary

The main contribution of this paper is to propose a mean-field Wasserstein gradient flow (MF-WGF) algorithm for implementing the MF variational inference and to build a general theoretical framework for analyzing its statistical and algorithmic convergence for a generic class of Bayesian models (under the frequentist perspective).

Methodology-wise, by viewing the KL divergence as an objective functional over the space of all factorized probability measures, we develop a minimization scheme for implementing the MF approximation based on a time-discretized WGF. For Bayesian models without latent variables, the proposed algorithm is a distributional version of parallel coordinate proximal descent for updating the constituting components in the MF approximation. For Bayesian latent variable models, the proposed algorithm resembles a distributional version of the classical Expectation–Maximization algorithm, consisting of an E-step of updating the latent variable variational distribution and an M-step of conducting steepest descent over the variational distribution of model parameters; the developed algorithm can also be viewed as an extension of the general Majorize-Minimization (MM) principal to minimizing a functional over the space of probability measures.

Theoretically, since a Wasserstein gradient flow extends the usual Euclidean gradient flow, we analogously define the notion of (local) "convexity" and "smoothness" for a generic functional in the Wasserstein space, under which (local) exponential convergence towards the optimum of the functional can be proved. To prove and quantify the algorithmic convergence, we illustrate how the "convexity" and "smoothness" of the objective functional in VI, which is the Kullback–Leibler divergence to the target posterior distribution, translate into conditions of the statistical model. As a result, we explicitly determine the algorithmic contraction rate in terms of various problem characteristics such as step size, sample size, smoothness of the likelihood function, missing data Fisher information, and observed data Fisher information. As an intermediate result in our proof, we show that the MF approximation to the posterior distribution inherits the consistency and contraction of the latter (Theorems 3 and 4); our result of a squared-exponential (or sub-Gaussian) type deviation bound on the MF approximation is stronger than most existing results that only implies a polynomially decay bound. In addition, unlike many previous works relying on case-by-case analysis (Bickel et al., 2013; Hall et al., 2011a,b; Ormerod and Wand, 2012; Titterington and Wang, 2006; Westling and McCormick, 2015; Zhang and Zhou, 2020) or applying some information inequality that relates the variational objective functional value

to certain risk function evaluating the estimation error Alquier and Ridgway (2020); Pati et al. (2018); Yang et al. (2020); Zhang and Gao (2020), our proof is general and based on identifying and analyzing the fixed point of the iterative scheme in MF-WGF. Our proof strategy offers a somewhat more direct insight explaining why MF approximation leads to consistent estimation, and can be potentially useful for investigating statistical properties of other approximation schemes beyond the mean-field.

Computation-wise, we discuss and compare two concrete numerical methods for realizing the JKO scheme. The first method is a Langevin SDE-based particle method for approximately realizing the JKO scheme, which is commonly used in the literature. However, according to our numerical experiments and discussion, the SDE approach suffers from a systematic error that remains undiminished even with more iterations and number of particles due to a long term bias term. This motivates us to propose an alternative method based on function approximation (FA) using neural networks. As we illustrate, the FA approach is unbiased, meaning that its output precisely solves the JKO scheme. Consequently, the unique fixed point of the iterative process from FA precisely yields the MF approximation solution; and there is no long term systematic bias arising from using a finite step size. We also highlight that different from the previous work on functional approximation such as Mokrov et al. (2021), our function approximation approach is based on an unconstrained formulation (c.f. Theorem 8) without the need of restricting the transport map into a gradient vector field of a convex potential. This property allows flexible choices of numerical methods for solving the corresponding optimization problem, and significantly enhances the convergence speed and overall performance of the algorithm.

1.3 Organization

The remainder of this paper is organized as follows. Section 2 provides some preliminary results and the problem formulation. Specifically, we start with some background introduction to optimal transport theory and Wasserstein gradient flows; then we provide some new theoretical results about contraction properties of a discretized Wasserstein gradient flow, called the one-step minimization movement or the JKO scheme, with an explicit contraction rate; lastly, we discuss the connection between Wasserstein gradient flows and mean-field variational inference, and formulate the problem to be addressed in this work. In Section 3, we first provide a general computational framework for mean-field inference via alternating minimization, and then propose a new algorithm based on the discretized Wasserstein gradient flow. Section 4 presents our main theoretical result about the statistical concentration of the mean-field approximation and the algorithmic contraction of the proposed algorithm. In Section 5, we introduce and compare two numerical methods, particle approximation via SDE and function approximation method, for implementing the JKO scheme. In Section 6, we apply our theoretical results to two representative examples, namely, the Gaussian mixture model and the mixture of regression model; we also conduct some numerical experiments to compliment the theoretical findings. All proofs and other technical details are postponed to an online supplementary material `https://arxiv.org/pdf/2207.08074`, which includes all the appendices.

1.4 Notation

We use $\mathscr{P}(\mathbb{R}^d)$ to denote the space of all probability measures on $\mathbb{R}^d$, and use $\mathscr{P}_2(\mathbb{R}^d)$ to denote the subset of $\mathscr{P}(\mathbb{R}^d)$ composed of all measures with finite second-order moment, i.e.

$$\mathscr{P}_2(\mathbb{R}^d) = \Big\{\mu \in \mathscr{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < \infty\Big\}.$$

Let $\mathscr{P}_2^r(\mathbb{R}^d)$ denote the space of all probability measures in $\mathscr{P}_2(\mathbb{R}^d)$ that admit a density function relative to the Lebesgue measure of $\mathbb{R}^d$. For any measure μ on $\mathbb{R}^d$ and map $T : \mathbb{R}^d \to \mathbb{R}^d$, the pushforward measure $\nu = T_{\#}\mu$ is defined as the unique measure on $\mathbb{R}^d$ such that $\nu(A) = \mu\big(T^{-1}(A)\big)$ holds for any measurable set A on $\mathbb{R}^d$. We use $D_{\mathrm{KL}}(p \,\|\, q)$ to denote the KL divergence between two probability measures p, $q \in \mathscr{P}_2^r(\mathbb{R}^d)$. Depending on the context, we may use upper letters to denote probability measures, and lower letters to denote their probability density functions. For any $\alpha \in [1, \infty)$, let $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be the function defined by $\psi_\alpha(x) = \exp(x^\alpha) - 1$. We use the notation $\|\xi\|_{\psi_\alpha} = \inf\big\{C > 0 : \mathbb{E}[\,\psi_\alpha(|\xi|/C)\,] \leqslant 1\big\}$ to denote the α-th order *Orlicz norm* of a real-valued random variable ξ (see Appendix G for a brief review). We also use $\mathcal{L}(\xi)$ to denote the law (distribution) of random variable ξ. We use $|||M|||_{\mathrm{op}} = \sup_{v \in \mathbb{S}^{n-1}} \|Mv\|$ to denote the matrix operator norm of a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, where $\mathbb{S}^{n-1}$ is the $(n-1)$-dimensional unit sphere. We use Id to denote the identity map.

2. Preliminary Results and Problem Formulation

In this section, we first briefly review some concepts and basic results from optimal transport theory. After that, we discuss the notion of Wasserstein gradient flow and its discrete-time version, and present some new results about the contraction of one-step discretized Wasserstein gradient flow, which will be useful in our later analysis of alternating minimization for solving mean-field variational inference. Finally, we setup the Bayesian framework, review the mean-field inference, and formulate the problem to be addressed in this paper. Further details and techniques, such as subdifferential calculus in the Wasserstein space for analyzing the optimization landscape of functionals of probability measures and its connection with the usual Gateaux derivative (a.k.a. first variation), are deferred to Appendix A.

2.1 Optimal transport and Wasserstein space

The Wasserstein space $\mathbb{W}_2(\mathbb{R}^d) = \big(\mathscr{P}_2(\mathbb{R}^d), W_2\big)$ is the separable metric space that endows $\mathscr{P}_2(\mathbb{R}^d)$ with the 2-Wasserstein metric W_2 (Ambrosio et al., 2008). In particular, the 2-Wasserstein distance between two distributions μ and ν in $\mathscr{P}_2(\mathbb{R}^d)$ is defined as

$$W_2^2(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \Big\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\gamma(x, y)\Big\}, \qquad \text{(KP)} \tag{1}$$

where $\Pi(\mu, \nu)$ consists of all possible distributions over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, and any $\gamma \in \Pi(\mu, \nu)$ is called a coupling between μ and ν. It can be proved (Section 5 of Santambrogio (2015)) that W_2 is indeed a metric on $\mathscr{P}_2(\mathbb{R}^d)$ and satisfies the triangle inequality; moreover, convergence with respect to W_2 is equivalent to the usual weak convergence of probability measures plus convergence of second moments. If one of the distributions, say

μ, is absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^d$, or $\mu \in \mathscr{P}_2^r(\mathbb{R}^d)$, then the optimal coupling $\gamma^* = (\mathrm{Id}, T^*)_{\#}\mu$ is unique (Theorem 1.22, Santambrogio (2015)) and supported on the graph of a map $T^* : \mathbb{R}^d \to \mathbb{R}^d$, called the optimal transport map from μ to ν; see Appendix A.1 for further properties of this optimal transport map.

2.2 Wasserstein gradient flow

Consider the problem of minimizing a functional $\mathcal{F} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$ in the Wasserstein space $\mathbb{W}_2(\mathbb{R}^d)$ via "steepest descent". A direct generalization of the ODE formulation of the Euclidean gradient flow (c.f. Appendix A.3) is to define a time-dependent measure $\rho_t \in \mathscr{P}_2^r(\mathbb{R}^d)$ satisfying $\partial_t \rho_t = -\nabla_{W_2}\mathcal{F}(\rho_t)$ for $t > 0$, with some initialization $\rho_0 \in \mathscr{P}_2^r(\mathbb{R}^d)$. Here $\nabla_{W_2}\mathcal{F}$ stands for some proper notion of gradient, or steepest (ascent) direction, of $\mathcal{F}$ with respect to the W_2 metric in $\mathbb{W}_2(\mathbb{R}^d)$. To formally define and prove the well-posedness of this Wasserstein gradient flow (WGF), one can first consider a minimization movement scheme, also called the Jordan-Kinderlehrer-Otto (JKO) scheme Jordan et al. (1998) (see Figure 1 for an illustration),

$$\rho_{k+1}^\tau = \operatorname*{argmin}_{\rho \in \mathscr{P}_2(\mathbb{R}^n)} \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho_k^\tau, \rho) \quad \text{for } k \geqslant 0; \qquad \text{(JKO)} \tag{2}$$

and then show by using a generalised version of Arzelà–Ascoli theorem that after suitable interpolation, the solution of this JKO scheme admits a limit as step size $\tau \to 0_+$; finally this limit as an absolutely continuous curve (proved by a priori estimate) in $\mathbb{W}_2(\mathbb{R}^d)$ is defined as the Wasserstein gradient flow for minimizing $\mathcal{F}$ starting from ρ_0. Details of a complete proof in the more general setting of gradient flows in metric spaces can be found in Chapter 3 of Ambrosio et al. (2008). A proof in the case of Fokker-Planck equation as the Wasserstein gradient flow of the KL functional (6) (c.f. Section 2.4) can be found in Jordan et al. (1998) or Chapter 8.3 of Santambrogio (2015).

Under the above perspective, it can be shown that the WGF for minimizing $\mathcal{F}$ (by taking the limit of JKO scheme as $\tau \to 0_+$) can be characterized by the following partial differential equation (PDE), also called continuity equation,

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t), \quad \text{with } v_t = -\nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t), \quad \text{for } t > 0, \tag{3}$$

where $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is the flow velocity vector field at time t, corresponding to the location-dependent steepest descent direction (i.e. negative subdifferential) given in Lemma A.1. Here, we have abused the notation by using $\rho_t \in \mathscr{P}_2^r(\mathbb{R}^d)$ to denote both the probability measure and its density function. Similarly, in the rest of the paper, we will use the notation ρ for a generic regular probability measure in $\mathscr{P}_2^r(\mathbb{R}^d)$ and its density. PDE (3) provides the Eulerian description of the WGF for functional $\mathcal{F}$, and motivates one numerical method for implementing WGF via functional approximation (Section 5).

In the continuity equation (3), we can view the time derivative $\partial_t \rho_t$ as the accumulation of probability mass, and intepret $\nabla \cdot (\rho_t v_t)$ as the "Wasserstein gradient" $\nabla_{W_2}\mathcal{F}$, where $\rho_t v_t$ is the flux and the divergence term $\nabla \cdot (\rho_t v_t)$ represents the difference in flow in versus flow out. Note that the continuity equation (3) may be interpreted as the equation governing the evolution of the density $\{\rho_t = (Y_t)_{\#}\rho_0 : t > 0\}$ of a family of particles initially distributed

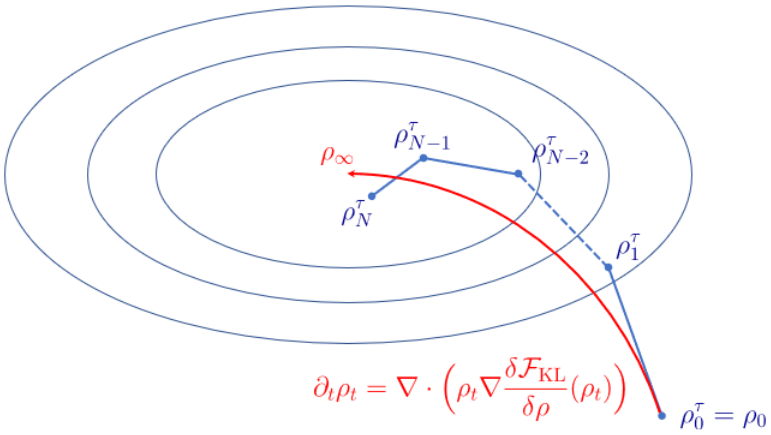


Figure 1: Illustration of a Wasserstein gradient flow (WGF, red curve) and its time-discretization via JKO-scheme (blue curve) with step size τ. For a functional $\mathcal{F}_{\mathrm{KL}}$ that is strictly convex along generalized geodesics, WGF converges to its global minimizer ρ_∞ exponentially fast. JKO-scheme discretizes the WGF, has the same limiting (or stationary) point ρ_∞ as WGF, and weakly converges to WGF as $\tau \to 0$.

according to ρ_0, and each of which follows the flow $\{Y_t : t > 0\}$. Here, the map $Y_t : \mathbb{R}^d \to \mathbb{R}^d$ is defined through $Y_t(x) = y_x(t)$ where, for any $x \in \mathbb{R}^d$, $\{y_x(t) : t \geqslant 0\}$ is the solution to the following ODE,

$$\dot{x}_t = v_t(x_t), \quad \text{for } t > 0, \quad \text{with } x_0 = x, \tag{4}$$

where v_t specifies the (steepest descent) direction of particles in the gradient flow. This ODE corresponds to a Lagrangian description of the WGF that characterizes the state of each individual "particle" at each time, rather than counting the number of "particles" sharing the same state (e.g., location and velocity), and motivates another numerical method for implementing WGF via particle approximation (Section 5).

2.3 Contraction of one-step minimization movement

The following functional $\mathcal{F}_{\tau,\mu} : \mathscr{P}_2^r(\mathbb{R}^d) \to (-\infty, \infty]$ defined as

$$\mathcal{F}_{\tau,\mu}(\nu) = \mathcal{F}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \mu), \tag{5}$$

has been used in defining the minimization movement scheme (2) for minimizing $\mathcal{F}$ on $\mathbb{W}_2(\mathbb{R}^d)$. We assume that for some $\tau_* > 0$, $\mathcal{F}_{\tau,\mu}$ admits at least a minimum point μ_τ, for all $\tau \in (0, \tau_*)$ and $\mu \in \mathscr{P}_2^r(\mathbb{R}^d)$. The map $\mu \mapsto \mu_\tau$ can be seen as a generalization from the usual Euclidean space to $\mathscr{P}_2^r(\mathbb{R}^d)$ of the proximal operator associated with functional $\tau\mathcal{F}$, where the Euclidean distance is replaced by the Wasserstein distance.

We can use μ_τ or $\mathcal{F}_{\tau,\mu}$ to define the one-step discretization of the Wasserstein gradient flow, which can then be used for both formally defining the gradient flow (as in Section 2.2) and providing a numeric scheme for approximating the gradient flow. Such a one-step discretization will also serve as the building block of the proposed MF-WGF with $\mathcal{F}$ being the KL divergence to the target posterior (c.f. Section 3.2). In the rest of this subsection, we provide a theoretical analysis of the one-step minimization movement of minimizing $\mathcal{F}_{\tau,\mu}$. This technical result will be useful in analyzing the convergence of the proposed MF-WGF method later.

Convexity plays an important role in proving convergence and deriving explicit convergence rates of gradient flows in Euclidean space. To extend the notion of convexity to the Wasserstein space, one approach is to consider convexity along generalized geodesics. This requires the target functional $\mathcal{F}$ to exhibit convexity along certain interpolating curve that connects any pair of probability measures in $\mathbb{W}_2(\mathbb{R}^d)$. For a formal definition and additional properties, please refer to Appendix A.4. For any $\lambda > 0$, we say that $\mathcal{F}$ is λ-convex along generalized geodesics if it is λ-convex along any generalized geodesic in the usual sense (as a univariate function under the constant speed parameterization of the curve). Using this notion, we have the following theorem about the contraction of one-step minimization movement. Its proof is left to Appendix B.3, which utilizes a key Lemma A.4 to derive a contraction with an explicit contraction factor.

Theorem 1. *Let $\mathcal{F} : \mathscr{P}_2^r(\mathbb{R}^d) \to (-\infty, \infty]$ be λ-convex along generalized geodesics. Then for any μ, $\pi \in \mathscr{P}_2^r(\mathbb{R}^d)$,*

$$(1 + \tau\lambda)\, W_2^2(\mu_\tau, \pi) \leqslant W_2^2(\mu, \pi) - 2\tau\big[\mathcal{F}(\mu_\tau) - \mathcal{F}(\pi)\big] - W_2^2(\mu_\tau, \mu),$$

where

$$\mu_\tau = \operatorname*{argmin}_{\rho \in \mathscr{P}_2^r(\mathbb{R}^d)} \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\mu, \rho).$$

In particular, if π^* is any minimizer of $\mathcal{F}$, then

$$W_2^2(\mu_\tau, \pi^*) \leqslant (1 + \tau\lambda)^{-1}\, W_2^2(\mu, \pi^*), \quad \forall \mu \in \mathscr{P}_2^r(\mathbb{R}^d).$$

As a direct consequence of the theorem, the time-discretized Wasserstein gradient flow for minimizing a λ-convex (along generalized geodesics) functional $\mathcal{F}$ obtained by repeatedly applying the one-step minimization movement achieves an exponential convergence to the unique global minimizer of $\mathcal{F}$, with contraction factor $(1 + \tau\lambda)^{-1} \in (0, 1)$ for any step size $\tau > 0$. Note that this convergence behavior is similar to the implicit Euler scheme for minimizing a λ-convex function on $\mathbb{R}^d$, while the explicit Euler scheme is convergent only when τ is smaller than some threshold inverse proportional to the largest eigenvalue of the Hessian $\nabla^2 F$, indicating the robustness and stability of implicit schemes.

2.4 KL divergence functional

In this paper, we are interested in functionals over $\mathscr{P}_2^r(\mathbb{R}^d)$ of the following form, due to the close connection with the KL divergence as the optimization objective in VI,

$$\mathcal{F}_{\mathrm{KL}}(\rho) = \underbrace{\int_{\mathbb{R}^d} V(x)\,\mathrm{d}\rho(x)}_{\text{potential energy } \mathcal{V}(\rho)} + \underbrace{\int_{\mathbb{R}^d} \log \rho(x)\,\mathrm{d}\rho(x)}_{\text{entropy } \mathcal{E}(\rho)}. \tag{6}$$

The KL functional $\mathcal{F}_{\mathrm{KL}}$ consists of an entropy functional $\rho \mapsto \int \log \rho \,\mathrm{d}\rho$ and a potential energy functional $\rho \mapsto \int V \mathrm{d}\rho$, where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential (function).

When specialized to the KL functional $\mathcal{F}_{\mathrm{KL}}$, the continuity equation (3) for characterizing its Wasserstein gradient flow becomes the famous Fokker-Planck equation

$$\frac{\partial \rho_t}{\partial t} - \Delta \rho_t - \nabla \cdot (\rho_t \nabla V) = 0, \tag{7}$$

since the first variation $\frac{\delta \mathcal{F}_{\mathrm{KL}}}{\delta \rho} = V + \log \rho + C$ (first variation is defined up to a constant) and $v_t = -\nabla V - \nabla \log \rho_t$. It is well known that the solution ρ_t to the Fokker-Planck equation also corresponds to the law of Langevin stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad X_0 \sim \rho_0. \tag{8}$$

This connection will motivate one of our discretizing schemes for realizing the Wasserstein gradient flow for $\mathcal{F}_{\mathrm{KL}}$ (c.f. Section H.1).

It turns out that the entropy $\mathcal{E}$ is convex along generalized geodesics (Proposition 9.3.9, Ambrosio et al., 2008) and the potential energy $\mathcal{V}$ is λ-convex along generalized geodesics if the corresponding potential function V is a λ-convex function over $\mathbb{R}^d$ (Proposition 9.3.2 Ambrosio et al., 2008). Therefore, using Theorem 1, we obtain the following corollary characterizing the contraction property of one-step movement minimization for minimizing the KL functional $\mathcal{F}_{\mathrm{KL}}$.

Corollary 2. *If potential $V : \mathbb{R}^d \to \mathbb{R}$ is a λ-convex function over $\mathbb{R}^d$, then for any μ, $\pi \in \mathscr{P}_2^r(\mathbb{R}^d)$,*

$$(1 + \tau\lambda)\, W_2^2(\mu_\tau, \pi) \leqslant W_2^2(\mu, \pi) - 2\tau\big[\mathcal{F}_{\mathrm{KL}}(\mu_\tau) - \mathcal{F}_{\mathrm{KL}}(\pi)\big] - W_2^2(\mu_\tau, \mu),$$

where

$$\mu_\tau = \operatorname*{argmin}_{\rho \in \mathscr{P}_2^r(\mathbb{R}^d)} \mathcal{F}_{\mathrm{KL}}(\rho) + \frac{1}{2\tau} W_2^2(\mu, \rho).$$

In particular, if $\pi^*(x) \propto e^{-V(x)}$ for $x \in \mathbb{R}^d$, then

$$W_2^2(\mu_\tau, \pi^*) \leqslant (1 + \tau\lambda)^{-1}\, W_2^2(\mu, \pi^*), \quad \forall \mu \in \mathscr{P}_2^r(\mathbb{R}^d).$$

2.5 Mean-field variational inference

A generic probabilistic model consists of a collection of observed variables $X \in \mathcal{X}$ and a collection of hidden variables $Z \in \mathcal{Z}$, where Z may contain model parameters and latent variables as its components in a Bayesian setting. The goal is to (approximately) learn the posterior distribution $p(Z \,|\, X) = p(X,\, Z)/p(X)$ of the hidden variables given the observed ones. In a typical problem setting, the joint distribution $p(X, Z)$ is only known up to a constant; therefore, the exact computation of $p(Z \,|\, X)$ is intractable due to the high-dimensional integral involved in computing the normalization constant.

A generic variational inference (VI) approaches this task by turning the integration problem into an optimization one as below,

$$\widehat{q} = \operatorname*{argmin}_{q \in \Gamma} \; D_{\mathrm{KL}}\big(q \,\|\, p(\cdot \,|\, X)\big), \tag{9}$$

where Γ is an user-specified distribution family over the hidden variable space $\mathcal{Z}$, called the variational family. In another word, VI uses a closest member (relative to KL divergence) in the variational family Γ to approximate the target posterior. The KL divergence is used as the discrepancy measure for two reasons: 1. it can be computed up to a constant without the knowledge of the normalization constant in the posterior; 2. it captures the information geometry in the statistical models.

MF inference is a special case of VI when the variational family Γ_{MF} is composed of all factorized q with the following form,

$$q(z) = q_1(z_1)\, q_2(z_2) \cdots q_m(z_m), \quad \text{for } z = (z_1,\, z_2,\, \ldots,\, z_m) \in \mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m,$$

where each component (block) z_j of z may contain more than one variables. In general, to alleviate the bias incurred by ignoring the dependence among the blocks $\{Z_j\}_{j=1}^m$, it is preferable to use a reduced number of blocks while maintaining the computational tractability of solving problem (9). In this work, we consider two model settings: Bayesian models with and without latent variables.

Bayesian models without latent variables. In this setting, the hidden variables Z solely consist of model parameters $\theta \in \Theta \subset \mathbb{R}^d$ and we consider mean-field approximation over (blocks of) components of θ. We consider a standard model setting where the observations $X = X^n = \{X_1, \cdots, X_n\}$ are i.i.d. given θ. We denote the prior and posterior distributions of θ as π_θ and π_n, respectively, where

$$\pi_n(\theta) = \frac{\pi_\theta(\theta) \prod_{i=1}^n p(X_i \,|\, \theta)}{\int_\Theta \pi_\theta(\theta) \prod_{i=1}^n p(X_i \,|\, \theta)\, \mathrm{d}\theta}, \quad \text{for } \theta \in \Theta. \tag{10}$$

We further divide the parameter space into m blocks, i.e., $\Theta = \bigotimes_{j=1}^m \Theta_j$, where $\Theta_j \subset \mathbb{R}^{d_j}$ and $d_1 + \cdots + d_m = d$. The corresponding MF approximation to π_n to be studied is

$$\widehat{q}_\theta = \bigotimes_{j=1}^d \widehat{q}_j \in \operatorname*{argmin}_{q = \otimes_{j=1}^d q_j} \; D_{\mathrm{KL}}(q \,\|\, \pi_n). \tag{11}$$

Our theoretical result in Section 4.1 demonstrates that point estimators obtained from above MF approximation achieve the same rate of convergence in estimation error as those

obtained from the full posterior π_n, under the frequentist perspective that assumes X^n to be generated from a true underlying data generating model indexed by a true parameter θ^*.

Bayesian models with latent variables. In this setting, we have observed variable $X = X^n$ as before but the hidden variable Z now includes model parameter $\theta \in \Theta \subset \mathbb{R}^d$ and a collection of latent variables $Z^n = \{Z_1, \cdots, Z_n\} \in \mathcal{Z}^n$, such that $(X_i, Z_i)_{i=1}^n$ are i.i.d. given θ. For simplicity, we assume the latent variables to be discrete, such as the latent class (cluster) indicators in Gaussian mixture models. Let π_θ denote the prior distribution defined on parameter space Θ. To maintain a minimal number of blocks for maximally reducing the potential bias, we consider the following (two-block) mean-field approximation over the parameter block θ and latent variables block Z^n,

$$(\widehat{q}_\theta,\, \widehat{q}_{Z^n}) = \underset{q_\theta \in \mathscr{P}(\Theta),\, q_{Z^n} \in \mathscr{P}(\mathcal{Z}^n)}{\operatorname{argmin}} D_{KL}(q_\theta \otimes q_{Z^n} \,\|\, \pi_n), \tag{12}$$

where in this case π_n denotes the joint posterior distribution of (θ, Z^n), given by

$$\pi_n(\theta, z^n) = \frac{\pi_\theta(\theta) \prod_{i=1}^n p(X_i, z_i \,|\, \theta)}{\sum_{z^n \in \mathcal{Z}^n} \int_\Theta \pi_\theta(\theta) \prod_{i=1}^n p(X_i, z_i \,|\, \theta)\, \mathrm{d}\theta}, \quad \text{for } \theta \in \Theta \ \text{ and } \ z^n \in \mathcal{Z}^n. \tag{13}$$

It is also possible to consider a full mean-field approximation by also factorizing q_θ over blocks of θ. However, this scheme may introduce additional complications without providing further insights due to its overlap with the first setting without latent variables. A similar theoretical result in Section 4.1 demonstrates the statistical optimality of point estimation using MF approximation (12).

Computation via coordinate ascent variational inference. Alternating minimization is a natural and commonly used algorithm for optimizing over quantities taking a product form as in MF inference. The idea of alternative minimizing is to optimize over one component of q at a time while fixing the others. Consider the generic MF approximation (9) and let $q_{-j}(z_{-j}) = \prod_{s \neq j} q_s(z_s)$ denote the joint distribution of z_{-j}, all components in z except for z_j. When optimizing over the j^{th} component q_j, one may explicitly solve the optimizer $q_j^* := \operatorname{argmin}_{q_j} D_{\mathrm{KL}}\big(q_j \otimes q_{-j} \,\|\, p(\cdot \,|\, X)\big)$ as

$$q_j^*(z_j) \propto \exp\Big\{ \int_{\mathcal{Z}_{-j}} \log p(z_j, z_{-j}, X)\, \mathrm{d}q_{-j}(z_{-j}) \Big\}, \quad \text{for } z_j \in \mathcal{Z}_j. \tag{14}$$

However, to make the computation of q_j^* tractable, one requires certain conditional conjugacy structures so that the integral inside the exponent can be explicitly calculated and the normalization constant of q_j^* can be identified. To avoid overly aggressive moves that may lead to non-convergence of the algorithm, it may be necessary to introduce a partial step size into the above update if q_j^* can be recognized as a member of some parametric family Bhattacharya et al. (2023), leading to the so-called coordinate ascent variational inference (CAVI) algorithm Bishop and Nasrabadi (2006).

Goal of this work. The main problem to be addressed in this work is to design a new class of computational algorithms for solving the above optimization problems for MF variational inference based on Wasserstein gradient flow while adapting the idea of alternating

minimization, and to study their theoretical properties. Since Wasserstein gradient flow directly operates over the space of probability measures, the new method does not need impose any extra restrictions on the MF variational family (which may unnecessarily increase the approximation error), and can be applied to Bayesian models without any structural constraint on the prior and data likelihood function. Moreover, a step size tuning parameter is naturally incorporated to prevent overly aggressive moves which can cause the algorithm to diverge.

3. Mean-Field Variational Inference via Wasserstein Gradient Flow

In this section, we propose a generic computational framework of MF variational inference for models with and without latent variables by alternating minimization and coordinate ascent in the Wasserstein space via repeatedly applying a one-step discretized Wasserstein gradient flow to components in the MF approximation.

3.1 Bayesian models without latent variables

Recall that the mean-field variational family $\Gamma = \{q = \bigotimes_{j=1}^m q_j : q_j \in \mathscr{P}_2(\Theta_j)\}$ is the set of all factorized distributions over m blocks of parameter θ. We use the shorthand $q_{-j}^{(k)} = \bigotimes_{l\neq j} q_l^{(k)}$ to denote the joint variational distribution of θ_{-j}, the parameter vector θ without its j-th block θ_j, in the k-th iteration. A standard algorithm for solving optimization problems involving multiple variables is alternating minimization. For technical convenience, we consider a parallel (simultaneous) update scheme for implementing the alternative minimization framework (14) to motivate our proposed method, which takes the following form under the current model setting,

$$q_j^{(k+1)} = \operatorname*{argmin}_{q_j} \; D_{\mathrm{KL}}(q_j \otimes q_{-j}^{(k)} \,\|\, \pi_n) \quad \text{for } j \in [m] \;\text{ and }\; k = 0, 1, \cdots . \tag{15}$$

Alternative minimization for solving MF can diverge due to its overly aggressive moves Bhattacharya et al. (2023). A common solution to avoid divergence when optimizing a multivariate function in Euclidean space is to use a one-step gradient descent, rather than fully minimizing the target function. In light of this, we propose replacing the update of q_j by solving (15) with a one-step discretized Wasserstein gradient flow for the functional $D_{\mathrm{KL}}(q_j \otimes q_{-j}^{(k)} \,\|\, \pi_n)$. This leads to a new computational framework for implementing the mean-field approximation (11) for Bayesian models without latent variables, which we call *mean-field Wasserstein gradient flow* (MF-WGF), by iteratively solving m sub-problems associated with discretized WGF in each iteration, which can be formulated as

$$q_j^{(k+1)} \in \operatorname*{argmin}_{q_j} \; D_{\mathrm{KL}}(q_j \otimes q_{-j}^{(k)} \,\|\, \pi_n) + \frac{1}{2\tau} W_2^2(q_j, q_j^{(k)}) \quad \text{for } j \in [m] \;\text{ and }\; k = 0, 1, \cdots . \tag{16}$$

The iterative updating formula (16) can be treated as the coordinate proximal descent algorithm in the Wasserstein space for minimizing the multi-input functional $D_{\mathrm{KL}}(q_1 \otimes \cdots \otimes q_m \,\|\, \pi_n)$. Here, we consider the parallel scheme which allows us to compute $q_j^{(k+1)}$ for different j parallelly, making the algorithm computationally efficient for large m.

Another appealing feature of MF-WGF is that the time discretization via the minimization movement scheme does not introduce any bias—the MF solution $\widehat{q}_\theta$ in (11) is the (unique) fixed point of the corresponding iterative procedure, as shown by our theoretical results in Section 4. Furthermore, the iterative procedure has exponential convergence to this solution. It is straightforward to show that $\widehat{q}_\theta = \bigotimes_{j=1}^m \widehat{q}_j$, as a fixed point to MF-WGF, satisfies the distributional equations

$$\widehat{q}_j(\theta_j) = \frac{\exp\big\{\int_{\Theta_{-j}} \log \pi_\theta(\theta) + \sum_{k=1}^n \log p(X_k \,|\, \theta)\, \mathrm{d}\widehat{q}_{-j}(\theta_{-j})\big\}}{\int_{\Theta_j} \exp\big\{\int_{\Theta_{-j}} \log \pi_\theta(\theta) + \sum_{k=1}^n \log p(X_k \,|\, \theta)\, \mathrm{d}\widehat{q}_{-j}(\theta_{-j})\big\}\, \mathrm{d}\theta_j}, \quad \text{for } j \in [m], \quad (17)$$

which can be proved by applying the first order optimality condition to (16) in terms of the first variation as described in Section A.2. Later, we will use this fixed point equation to show the concentration of MF approximation $\widehat{q}_\theta$ towards the true parameter θ^* (c.f. Theorem 3, also see Section 4.1 for a sketched proof). Unlike Bayesian latent variable models, we do not need this concentration property to prove the linear convergence of $\widehat{q}^{(k)}$ towards $\widehat{q}_\theta$ in the sense of W_2 metric, as stated in Theorem 5 in the next section.

3.2 Bayesian latent variable models

Due to the conditional independence among discrete latent variables $Z_1, \ldots, Z_n$ given θ and X^n, it is easy to verify that any minimizer $\widehat{q}_{Z^n}$ of optimization problem (12) also factorizes as $\widehat{q}_{Z^n} = \bigotimes_{i=1}^n \widehat{q}_{Z_i}$. Consequently, the alternating minimization framework (14) for solving mean-field optimization (12) under this model setting can be formulated as: for iteration $k = 0, 1, \ldots,$

$$\textbf{Latent variable update:} \quad q_{Z_i}^{(k+1)} = \Phi(q_\theta^{(k)}, X_i), \quad i = 1, 2, \ldots, n, \quad \text{with}$$

$$\Phi(q_\theta, X_i)(z) = \frac{\exp\big\{\mathbb{E}_{q_\theta} \log p(z \,|\, X_i, \theta)\big\}}{\sum_{z \in \mathcal{Z}} \exp\big\{\mathbb{E}_{q_\theta} \log p(z \,|\, X_i, \theta)\big\}}, \quad z \in \mathcal{Z}; \quad (18)$$

$$\textbf{Parameter update:} \quad q_\theta^{(k+1)} = \operatorname*{argmin}_{q_\theta} \; V_n\big(q_\theta \,\big|\, q_\theta^{(k)}\big), \quad \text{with}$$

$$\text{(sample energy functional)} \quad V_n(q_\theta \,|\, q'_\theta) := n\, \mathbb{E}_{q_\theta}\big[U_n(\theta;\, q'_\theta)\big] + D_{\mathrm{KL}}(q_\theta \,||\, \pi_\theta), \quad \text{and}$$

$$\text{(sample potential function)} \quad U_n(\theta,\, q'_\theta) := -\frac{1}{n} \sum_{i=1}^n \sum_{z \in \mathcal{Z}} \log p(X_i, z \,|\, \theta)\, \Phi(q'_\theta, X_i)(z),$$

where $\Phi : \mathscr{P}^r(\Theta) \times \mathcal{X} \mapsto \mathscr{P}(\mathcal{Z})$ denotes the map that turns a (q_θ, x) pair to a probability measure (pmf) over $\mathcal{Z}$. Since $\mathcal{Z}$ is discrete, the latent variable update can be easily performed using the closed form formula. The V_n functional above resembles the Q function computed in the E-step of a generic EM algorithm, and is equivalent up to a constant to the KL divergence $D_{KL}\big(q_\theta \otimes q_{Z^n}^{(k+1)} \,\big\|\, \pi_n\big)$.

Writing the updating formula for q_θ via minimizing the KL divergence functional V_n is more convenient for the design of algorithms and theoretical analysis. Since Z_i's are discrete, in the preceding alternating minimization algorithm, updating q_{Z^n} with a given q_θ amounts to solving $\min_{q_{Z^n}} D_{KL}(q_\theta \otimes q_{Z^n} \,\|\, \pi_n)$, which admits a closed form expression with tractable normalization. However, the step of updating q_θ by solving for the exact

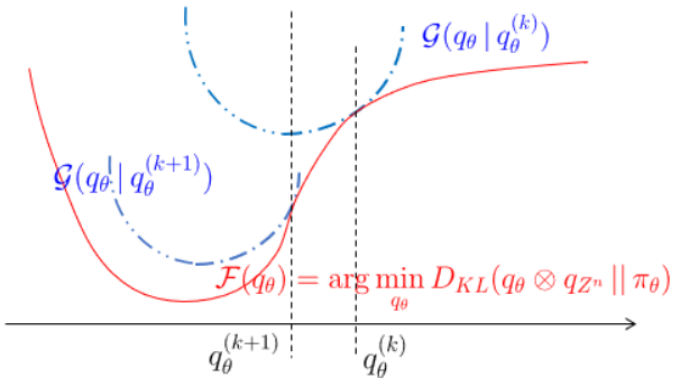


Figure 2: Mean-field Wasserstein gradient flow (MF-WGF) as an extension of the Majorize-Minimization (MM) algorithm Lange (2016) for minimizing (profile-KL) functional $\mathcal{F} = \min_{q_{Z^n}} D_{KL}\big(q_\theta \otimes q_{Z^n} \,\|\, \pi_n\big)$ over the space of all probability measures on parameter space Θ. Here, $\mathcal{G}(q_\theta \,|\, q'_\theta) := V_n(q_\theta \,|\, q'_\theta) + \frac{1}{2\tau} W_2^2(q_\theta,\, q'_\theta)$ majorizes $\mathcal{F}$.

minimizer of $V_n(\cdot \,|\, q_\theta^{(k)})$ may not be computationally tractable unless some conditional conjugacy condition is satisfied. Instead, we view $V_n(\cdot \,|\, q_\theta^{(k)})$ as the KL divergence functional over $\mathscr{P}_2^r(\Theta)$ and propose to update q_θ via its associated one-step discretized Wasserstein gradient flow. Note that in the situation where Z^n is continuous, we may also apply a one-step discreteized Wasserstein gradient flow to update q_{Z^n} rather than exactly minimize $D_{KL}(q_\theta \otimes q_{Z^n} \,\|\, \pi_n)$ over q_{Z^b}; the resulting algorithm then becomes coordinate descent over the space of all factorized probability distributions. We leave the formal study of this case to future work.

The perspective of viewing the parameter update step as minimizing a time-dependent KL divergence functional over $\mathscr{P}_2(\Theta)$ leads to our new computational framework of MF-WGF for Bayesian latent variable models. More precisely, MF-WGF involves iteratively cycling through the following two steps for iteration $k = 0, 1, \ldots$:

Step 1 *(Local latent variable):* For $i = 1, \ldots, n$, compute $q_{Z_i}^{(k+1)}$ based on the updating formula (18);

Step 2 *(Global model parameter):* Compute the energy functional $V_n(q_\theta \,|\, q_\theta^{(k)})$ using the most recent $q_{Z^n}^{(k+1)} = \bigotimes_{i=1}^n q_{Z_i}^{(k+1)}$, and update q_θ via the one-step minimization movement scheme (2) with objective functional $V_n(q_\theta \,|\, q_\theta^{(k)})$,

$$q_\theta^{(k+1)} = \operatorname*{argmin}_{q_\theta} \; V_n(q_\theta \,|\, q_\theta^{(k)}) + \frac{1}{2\tau} W_2^2(q_\theta,\, q_\theta^{(k)}). \tag{19}$$

The two steps of MF-WGF resemble a distributional version of the E-step and the M-step respectively in the classical EM algorithm for dealing with missing data problems. One can also view MF-WGF as an Majorize-Minimization (MM) algorithm Lange (2016) for distributional optimization (see Figure 2 for an illustration) where $\mathcal{G}(q_\theta \,|\, q'_\theta) := V_n(q_\theta \,|\, q'_\theta) + \frac{1}{2\tau} W_2^2(q_\theta,\, q'_\theta)$ serves as the majorized version of the (profile) objective functional $q_\theta \mapsto$

$\min_{q_{Z^n}} D_{KL}\big(q_\theta \otimes q_{Z^n} \,\big\|\, \pi_n\big)$, where q_{Z^n} has been profiled out since the minimizing over q_{Z^n} admits a closed form solution as in Step 1 of MF-WGF.

Similar to Bayesian models without latent variables, in Section 4 we show that the MF solution $(\widehat{q}_\theta, \widehat{q}_{Z^n})$ in (12) is a unique fixed point of the corresponding iterative procedure in a constant radius W_2-neighhorhood around the solution, and the iterative procedure has exponential convergence to this solution given it is initialized in this neighborhood. It is straightforward to show that $\widehat{q}_\theta$, as a fixed point to MF-WGF, satisfies

$$\mu(\theta) = \frac{1}{Z_n(\mu)}\, \pi_\theta(\theta)\, e^{-n\, U_n(\theta;\mu)}, \quad \text{with } Z_n(\mu) = \int_\Theta \pi_\theta(\theta)\, e^{-n\, U_n(\theta;\mu)}\, \mathrm{d}\theta, \tag{20}$$

which can be proved by applying the first order optimality condition to (19) in terms of the first variation (see Appendix D.2 for further details). This fixed point equation is helpful to show the concentration of MF approximation $\widehat{q}_\theta$ towards the true parameter θ^* (c.f. Theorem 4, also see Section 4.1 for a sketched proof). Heuristically, when n is large, $\widehat{q}_\theta$ is expected to concentrate around the point mass measure $\delta_{\theta*}$ at θ^*, so that we can roughly approximate $\widehat{q}_\theta$ by the right hand side of (20) with μ being replaced by $\delta_{\theta*}$; then the convergence follows by the fact that θ^* approximately minimizes the potential $U_n(\theta;\, \delta_{\theta*})$.

4. Theoretical Results

In this subsection, we present two main theoretical results of this work: concentration of the MF approximation $\widehat{q}_\theta$ to the true parameter θ^*, and the convergence of the proposed MF-WGF algorithm. In the next section, we will apply the theoretical results to three representative examples by verifying the assumptions. All proofs are deferred to the Appendices in the supplement of the paper.

4.1 Analysis of mean-field approximation

We adopt the frequentist perspective by assuming that data X^n are generated from a data generating model indexed by a true parameter θ^*. Before presenting the formal result, we make the following assumptions, most are standard for proving concentration of Bayesian posteriors Ghosal et al. (2000); Shen and Wasserman (2001) and their MF counterpart Alquier and Ridgway (2020); Pati et al. (2018); Yang et al. (2020); Zhang and Gao (2020).

Assumption A.1 (test condition)**.** *For some constants $c_1, c_2 > 0$ and any $\varepsilon > c_1\sqrt{\log n/n}$, there is a test function ϕ_n, such that*

$$\mathbb{E}_{\theta*}[\phi_n] \leqslant e^{-c_2 n\varepsilon^2}, \qquad \sup_{\theta:\exists j\in[m], s.t. \|\theta_j-\theta_j^*\|>\varepsilon} \mathbb{E}_\theta[1-\phi_n] \leqslant e^{-c_2 n\varepsilon^2}.$$

In a typical parametric setting, the existence of such a test can be proved by decomposing $\{\theta_j : \|\theta_j - \theta_j^*\| > \varepsilon\}$ into a countable union of annuluses. Each annuluses can be covered by a finite number of balls, within each ball the likelihood ratio type test can be employed Birgé (1979); Le Cam (2012). When $m = 1$ in Bayesian latent variable models, this is just the standard test condition discussed in (Ghosal et al., 2000, Section 7).

Assumption A.2 (prior thickness). *There is a measure $\widetilde{Q} = \otimes_{j=1}^m \widetilde{Q}_j$, subsets $\widetilde{\Theta}_j \subset \Theta_j$ for $j \in [m]$, and positive constants c_3 and c_4, such that for any $\theta \in \widetilde{\Theta} := \otimes_{j=1}^m \widetilde{\Theta}_j$ we have*

$$D_{\mathrm{KL}}\big(p(\cdot \,|\, \theta^*) \,\|\, p(\cdot \,|\, \theta)\big) \leqslant c_4 \varepsilon_n^2, \quad \int_{\mathcal{X}} \Big(\log \frac{p(x \,|\, \theta^*)}{p(x \,|\, \theta)}\Big)^2 p(x \,|\, \theta^*)\, \mathrm{d}x \leqslant c_4 \varepsilon_n^2,$$
$$\log \frac{\mathrm{d}\widetilde{Q}}{\mathrm{d}\Pi_\theta} \leqslant c_4 n \varepsilon_n^2 \quad \textit{and} \quad \log \widetilde{Q}(\widetilde{\Theta}) = \sum_{j=1}^m \log \widetilde{Q}_j(\widetilde{\Theta}_j) \geqslant -c_3 n \varepsilon_n^2,$$

where Π_θ denotes the prior distribution, and $\varepsilon_n = M\sqrt{\log n / n}$ for some $M > 1$.

The case of $m = 1$ reduces to the Bayesian posterior without MF approximation. Under $m = 1$, this assumption is implied by the standard prior thickness assumption Ghosal et al. (2000) by taking $\widetilde{Q} = \Pi_\theta$. When $m > 1$, this assumption requires the existence of a fully factorized probability measure $\widetilde{Q}$ in the MF family that is close to the prior distribution and puts enough mass around the ground truth θ^*.

For Bayesian latent variable models, we need an additional assumption on the conditional likelihood function of latent variable Z given θ and observation X.

Assumption A.3 (local bound of KL divergence). *The marginal distribution $p(x \,|\, \theta)$ of observation X under θ and the conditional distribution $p(z \,|\, x, \theta)$ of latent variable Z given $X = x$ and θ satisfy*

$$D_{KL}\big(p(\cdot \,|\, x, \theta^*) \,\|\, p(\cdot \,|\, x, \theta)\big) \leqslant G(x)\, \varepsilon_n^2, \quad \forall\, \theta \in \widetilde{\Theta}, \tag{21}$$

where $\widetilde{\Theta}$ is the local neighborhood of θ^* defined in Assumption A.2, and $G(X)$ is a sub-exponential random variable with parameters σ_4 under $p(\cdot \,|\, \theta^*)$, i.e. $\mathbb{E}_{\theta^*}\big[\exp\big\{\sigma_4^{-1}|G(X)|\big\}\big] \leqslant 2$.

This assumption is a mild condition. In fact, we could expect

$$\varepsilon_n^2 \gtrsim D_{\mathrm{KL}}\big(p(\cdot \,|\, \theta^*) \,\|\, p(\cdot \,|\, \theta)\big) \geqslant D_H^2\big(p(\cdot \,|\, \theta^*), p(\cdot \,|\, \theta)\big) \gtrsim \|\theta - \theta^*\|^2,$$

where D_H represents the Hellinger distance. Here the first inequality is due to Assumption A.2 and $\widetilde{\Theta} \subset \Theta$; the last inequality usually holds when Θ is compact (Section 5, Ghosal et al., 2000). Therefore, Assumption A.3 is implied by a quadratic growth of KL divergence for $\theta \in \widetilde{\Theta}$, i.e. $D_{\mathrm{KL}}(p(\cdot \,|\, x, \theta^*) \,\|\, p(\cdot \,|\, x, \theta)) \lesssim G(x)\|\theta - \theta^*\|^2$. This quadratic growth property holds if the logarithms of both distributions (density or mass function) are twice differentiable with controlled Hessians.

For simple presentation, we adopt the assumption that $G(X)$ is sub-exponential to derive a high probability upper bound of $n^{-1}\sum_{i=1}^n G(X_i)$. This sub-exponential assumption on $G(X)$ can be generalized to $G(X)$ having a finite Orlicz-norm. See Appendix G for the definition of the Orlicz norm of a random variable and further details.

Bayesian models without latent variables. Recall that $\widehat{q}_\theta$ is the solution of the mean-field optimization problem (11), which should satisfy the following optimality condition (see Lemma C.1 in Appendix C.1),

$$\widehat{q}_j(\theta_j) = \frac{\exp\big\{\int_{\Theta_{-j}} \log \pi_\theta(\theta) + \sum_{k=1}^n \log p(X_k \,|\, \theta)\, \mathrm{d}\widehat{q}_{-j}(\theta_{-j})\big\}}{\int_{\Theta_j} \exp\big\{\int_{\Theta_{-j}} \log \pi_\theta(\theta) + \sum_{k=1}^n \log p(X_k \,|\, \theta)\, \mathrm{d}\widehat{q}_{-j}(\theta_{-j})\big\}\, \mathrm{d}\theta_j}, \quad \text{for } j \in [m]. \tag{22}$$

A major challenge in our analysis of the MF solution $\widehat{q}_\theta$ is how to deal with the normalization constant (denominator) in the preceding display, which depends on $\widehat{q}_\theta$ and complicates the analysis. To address this issue, we rewrite equation (22) by adding a θ_j-independent term to both nominator and denominator,

$$\widehat{q}_j(\theta_j) = \frac{\exp\big\{\int_{\Theta_{-j}} \log \frac{\pi_\theta(\theta)}{\widetilde{Q}_j(\theta_j)\widehat{q}_{-j}(\theta_{-j})} + \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta)}{p(X_i \,|\, \theta^*)} \,\mathrm{d}\widehat{q}_{-j}\big\}\widetilde{Q}_j(\theta_j)}{\int_{\Theta_j} \exp\big\{\int_{\Theta_{-j}} \log \frac{\pi_\theta(\theta)}{\widetilde{Q}_j(\theta_j)\widehat{q}_{-j}(\theta_{-j})} + \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta)}{p(X_i \,|\, \theta^*)} \,\mathrm{d}\widehat{q}_{-j}\big\} \,\mathrm{d}\widetilde{Q}_j}.$$

To prove the concentration of $\widehat{q}_j$ around the parameter θ_j^*, we can proceed as the usual steps for proving posterior concentration (e.g. Ghosal et al. (2000); Shen and Wasserman (2001)) by proving an upper and a lower bound to the numerator and the denominator respectively. Applying a union bound then yields the concentration of $\widehat{q}_\theta$ around θ^* due to the factorization structure of $\widehat{q}_\theta$.

For the lower bound to the denominator in the preceding display, denoted as D_j, we utilize the following equivalent expression,

$$\log D_j = -\widetilde{W}_n(\widehat{q}_\theta) = -\min_{q_\theta = \bigotimes_{j=1}^m q_j} \widetilde{W}_n(q_\theta), \quad \text{with} \tag{23}$$

$$\widetilde{W}_n(q_\theta) = \int_\Theta \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta^*)}{p(X_i \,|\, \theta)} \,\mathrm{d}q_1(\theta_1) \cdots \mathrm{d}q_m(\theta_m) + D_{\mathrm{KL}}(q_1 \otimes \cdots \otimes q_m \,\|\, \pi_\theta).$$

Here, functional $\widetilde{W}_n$ is, up to a q_θ-independent constant, the same as the objective functional $q_\theta = \bigotimes_{j=1}^m q_j \mapsto D_{KL}\big(q_\theta \,\|\, \pi_n\big)$. Thus, $\widehat{q}_\theta$ minimizes $\widetilde{W}_n$. Since $\widehat{q}_\theta$ is expected to be concentrated around θ^*, we may use $-\widetilde{W}_n(q_\theta)$ with some carefully constructed q_θ (the $\widetilde{Q}$ from Assumption A.2) suitably concentrated around θ^* (e.g. a uniform distribution supported on a small neighborhood around θ^*) for providing a lower bound to $\log D_j$.

For the upper bound to the numerator, the first term can be controlled by directly applying Jensen's inequality; the second term $\sum_{i=1}^n \log \frac{p(X_i \,|\, \theta)}{p(X_i \,|\, \theta^*)}$, which is roughly negative n times $D_{\mathrm{KL}}\big[\, p(\cdot \,|\, \theta^*) \,\big\|\, p(\cdot \,|\, \theta)\big]$ since $\{X_i\}_{i=1}^n$ are marginally i.i.d. from $p(\cdot \,|\, \theta^*)$ under the frequentist perspective. To formally bound the numerator, or more precisely, the integral of numerator over set $\Theta_\varepsilon = \{\|\theta - \theta^*\| \geqslant \varepsilon\}$ for suitably large $\varepsilon > 0$, we use the commonly adopted test condition (i.e., Assumption A.1) for uniformly controlling the log-likelihood ratio process over Θ_ε.

The following theorem shows the concentration of $\widehat{q}_\theta$ by characterizing the tail probability of being away from the true parameter θ^*. Recall that we are adopting a frequentist perspective, where the randomness in all high probability bound is coming from the randomness in the samples $X_1, \cdots, X_n$ that are generated under a true parameter θ^*.

Theorem 3 (Exponential posterior concentration without latent variables)**.** *Under Assumptions A.1 and A.2, if the sample size satisfies $c_4 n^{c_2 M^2} \geqslant 3mM^2 \log n$, then for any $M \geqslant 1$, the MF variational approximation $\widehat{Q}_\theta$ to the posterior distribution of θ satisfies the following with probability at least $1 - \frac{2c_4}{n\varepsilon_n^2} = 1 - \frac{2c_4}{M^2 \log n}$,*

$$\widehat{Q}_\theta\big(\exists\, j \in [m] \text{ s.t. } \|\theta_j - \theta_j^*\| > \varepsilon\big) \leqslant e^{-c_2 n \varepsilon^2/2},$$
$$\text{for all} \quad \varepsilon > M\Big(3 + c_1 + \frac{2c_4 + c_3 + 1}{c_2}\Big)\sqrt{\frac{\log n}{n}}. \tag{24}$$

Bayesian latent variable models. Let $\mathcal{Z} = \{1, 2, \ldots, K\}$ be the support of each discrete latent variable. In this case, $\widehat{q}_\theta$ is the solution of the mean-field optimization problem (12), satisfying (see Lemma C.3 in Appendix C.2 and Appendix D.2),

$$\widehat{q}_\theta(\theta) = \frac{1}{\widehat{Z}_n}\, \pi_\theta(\theta)\, e^{-n\, U_n(\theta,\, \widehat{q}_\theta)}, \quad \text{with } \widehat{Z}_n = \int_\Theta \pi_\theta(\theta)\, e^{-n\, U_n(\theta,\, \widehat{q}_\theta)}\, \mathrm{d}\theta, \tag{25}$$

where the (sample) potential function $U_n : \Theta \times \mathscr{P}^r(\Theta) \to \mathbb{R}$ is

$$U_n(\theta,\, q_\theta) = -\frac{1}{n} \sum_{i=1}^n \sum_{z=1}^K \log p(X_i, z \,|\, \theta)\, \Phi(q_\theta, X_i)(z), \quad \text{where} \tag{26}$$

$$\Phi(\widehat{q}_\theta, X_i)(z) := \widehat{q}_{Z_i}(z) = \frac{\exp\big\{\mathbb{E}_{\widehat{q}_\theta}\big[\log p(X_i, z \,|\, \theta)\big]\big\}}{\sum_{k=1}^K \exp\big\{\mathbb{E}_{\widehat{q}_\theta}\big[\log p(X_i, k \,|\, \theta)\big]\big\}}, \quad z \in [K].$$

Similar to the analysis of Bayesian models without latent variables, we rewrite equation (25) by adding a θ-independent term to both nominator and denominator,

$$\widehat{q}_\theta(\theta) = \frac{\exp\big\{ -\sum_{i=1}^n \sum_{z=1}^K \Phi(\widehat{q}_\theta, X_i)(z) \log \frac{\Phi(\widehat{q}_\theta, X_i)(z)}{p(z \,|\, X_i, \theta)} - \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta^*)}{p(X_i \,|\, \theta)} \big\}\, \pi_\theta(\theta)}{\int_\Theta \exp\big\{ -\sum_{i=1}^n \sum_{z=1}^K \Phi(\widehat{q}_\theta, X_i)(z) \log \frac{\Phi(\widehat{q}_\theta, X_i)(z)}{p(z \,|\, X_i, \theta)} - \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta^*)}{p(X_i \,|\, \theta)} \big\}\, \mathrm{d}\pi_\theta(\theta)}.$$

For the lower bound to the denominator, denoted as D_n, we can use the following equivalent expression, which is more convenient to analyze,

$$\log D_n = -W_n(\widehat{q}_\theta) = -\min_{q_\theta} W_n(q_\theta), \quad \text{with } W_n(q_\theta) = \tag{27}$$

$$\int_\Theta \bigg\{ \sum_{i=1}^n \sum_{z=1}^K \Phi(q_\theta, X_i)(z) \log \frac{\Phi(q_\theta, X_i)(z)}{p(X_i, z \,|\, \theta)} + \sum_{i=1}^n \log \frac{p(X_i \,|\, \theta^*)}{p(X_i \,|\, \theta)} \bigg\}\, \mathrm{d}q_\theta(\theta) + D_{\mathrm{KL}}(q_\theta \,\|\, \pi_\theta).$$

Here, functional W_n is, up to a q_θ-independent constant, the same as the (profile) objective functional $q_\theta \mapsto \min_{q_{Z^n}} D_{KL}\big(q_\theta \otimes q_{Z^n} \,\big\|\, \pi_n\big)$ after q_{Z^n} being maxed out or replaced by $\Phi(q_\theta, X_i)$; so $\widehat{q}_\theta$ minimizes W_n. Again, we may use $-W_n(q_\theta)$ with some carefully constructed q_θ suitably concentrated around θ^* (e.g. a uniform distribution supported on a small neighborhood around θ^*) for providing a lower bound to $\log D_n$.

For the upper bound to the numerator, the second term can be treated in the same way as in Bayesian models without latent variables. For the first term in the exponent, just note that

$$-\sum_{z=1}^K \Phi(\widehat{q}_\theta, X_i)(z) \log \frac{\Phi(\widehat{q}_\theta, X_i)(z)}{p(z \,|\, X_i, \theta)}$$

is the negative KL divergence between two discrete measures, and therefore is non-positive.

The following theorem shows the concentration of $\widehat{q}_\theta$ by characterizing the tail probability of being away from the true parameter θ^* in Bayesian latent variable models.

Theorem 4 (Exponential posterior concentration with latent variables)**.** *Under Assumptions A.1, A.2, and A.3, if the sample size satisfies* $6M \log n \leqslant \min\{n^{c_2 M}, e^{4n\sigma_4^{-1}}\}$*, then for*

any $M \geqslant 1$, the MF variational approximation $\widehat{Q}_\theta$ to the marginal posterior of θ satisfies the following with probability at least $1 - \frac{2c_4}{n\varepsilon_n^2} = 1 - \frac{2c_4}{M^2 \log n}$,

$$\widehat{Q}_\theta\big(\|\theta - \theta^*\| > \varepsilon\big) \leqslant e^{-c_2 n\varepsilon^2/2}, \tag{28}$$

$$\text{for all} \quad \varepsilon \geqslant M\Big(3 + c_1 + \frac{\mathbb{E}[G(X)] + c_3 + c_4 + 2}{c_2}\Big)\sqrt{\frac{\log n}{n}}. \tag{29}$$

Concentration properties of variational inference have been studied in the recent literature under different criteria. Alquier and Ridgway (2020) and Yang et al. (2020) consider a variant of the usual variational inference, called the α-variational inference, obtained by raising the likelihood to a fractional power $\alpha \in (0, 1]$ to facilitate the theoretical analysis. They prove upper bounds for the variational Bayes risk, defined as the expected Rényi divergence with respect to their α-fractional variational posterior. When α is strictly small than one, they only need a prior concentration assumption (similar to our Assumption A.2); however, under some mild conditions their risk function behaves like the second moment $\mathbb{E}_{\widehat{Q}_\theta}\big[\|\theta - \theta^*\|^2\big]$, which is much weaker than our sub-Gaussian type tail result. For the usual variational inference (or α-variational inference with $\alpha = 1$), Pati et al. (2018); Yang et al. (2020); Zhang and Gao (2020) proves high probability upper bounds to some similar variational Bayes risks that scales as $\mathbb{E}_{\widehat{Q}_\theta}\big[\|\theta - \theta^*\|^2\big]$ under similar test conditions (as our Assumption A.1) and a stronger version of the prior concentration assumption. Their proofs avoid assuming the compactness of parameter space by considering a sequence of sieve sets. Han and Yang (2019) proves a similar sub-Gaussian concentration result as ours; their proof is based on a perturbation analysis specifically tailored to the MF approximation and does not seem easily generalizable to other variatioal or hybrid schemes.

Our proof technique is very different from existing proofs of the variational posterior concentration in the literature, most of which are based on applying the variational characterization of KL divergence, $D_{\mathrm{KL}}(p \,||\, q) = \sup_h \big\{ \int hp - \log(\int e^h q)\big\}$. We instead view the variational posterior $\widehat{Q}_\theta$ as a point in the Wasserstein space $\mathbb{W}_2(\mathbb{R}^d)$ that minimizes a KL divergence functional, and uses its first order optimality condition (or equivalently, its stationarity to the time-discrete WGF) to show the concentration via subdiffential calculus in $\mathbb{W}_2(\mathbb{R}^d)$. This general perspective might be useful in extending the developed proof technique to other approximation scheme beyond MF, such as many recent generative model based variational inference procedures. Our obtained rate of convergence is also nearly optimal in the parametric setting; and the assumptions we made needed are standard in Bayesian asymptotics literature, and appears to be among the weakest in the context of variational inference. In addition, our proof techniques can be straightforwardly extended to non-parametric settings where the optimal rate of convergence is slower than root-n.

4.2 Analysis of MF-WGF algorithm

We separately analyze the convergence of MF-WGF for Bayesian models with/without latent variables.

Bayesian models without latent variables. Recall that in the k-th iteration, the MF-WGF algorithm updates the joint variational distribution $\bigotimes_{j=1}^m q_j^{(k)}$ into $\bigotimes_{j=1}^m q_j^{(k+1)}$ with

$$q_j^{(k+1)} = \operatorname*{argmin}_{q_j} \; n\,\mathbb{E}_{q_j \otimes q_{-j}^{(k)}}\big[U_n\big] + D_{\mathrm{KL}}\big(q_j \otimes q_{-j}^{(k)} \,\|\, \pi_\theta\big) + \frac{1}{2\tau} W_2^2(q_j, q_j^{(k)}), \quad j \in [m],$$

where $U_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p(X_i \,|\, \theta)$ is the sample potential function. The corresponding population-level potential function is

$$U(\theta) = -\int_{\mathbb{R}^d} \log p(x \,|\, \theta)\, p(\mathrm{d}x \,|\, \theta^*). \tag{30}$$

We expect that U_n and U are uniformly close enough when the sample size n is sufficiently large (e.g. Theorem 1 in Mei et al. (2018)).

Before formally presenting our result, we begin by introducing several assumptions that are commonly used to prove exponential convergence of iterative algorithms in optimization literature.

Assumption B.1 (strong convexity of population-level potential)**.** *There exists $\lambda > 0$ such that U is λ-strongly convex, i.e.*

$$U\big((1-t)\theta + t\theta'\big) \leqslant (1-t)\,U(\theta) + t\,U(\theta') - \frac{\lambda}{2}\,t(1-t)\|\theta - \theta'\|^2$$

for all $t \in [0,1]$ and $\theta, \theta' \in \Theta$. Moreover, the parameter space $\Theta = \bigotimes_{j=1}^m \Theta_j \subset \mathbb{R}^d$ is convex and contained a ball centered at the origin with radius R.

By the definition of $U(\theta)$ in Equation (30), a sufficient condition for Assumption B.1 is the strong log-concavity of the likelihood function $p(x \,|\, \theta)$, which is usually made in the literature of sampling and optimization on the space of probability distributions (Lee et al., 2021; Salim et al., 2020; Wibisono, 2018). This strong convexity assumption guarantees that $D_{\mathrm{KL}}(\cdot \,\|\, \pi_n)$ is strongly convex along generalized geodesics on $\mathscr{P}_2^r(\Theta_1) \times \cdots \times \mathscr{P}_2^r(\Theta_m) \subset \mathscr{P}_2^r(\Theta)$. It is possible to relax this global strong convexity to a local strong convexity within a small but constant-radius neighborhood around the true parameter θ^*, which is always true for regular models with non-singular Fisher information matrix. One simple strategy is to assume that both the prior and the initialization distribution of the algorithm are supported within this neighborhood. In practice, one can construct this initialization distribution by identifying a reasonably good initial point estimate of θ^*, for example, using simple and fast methods such as the method of moments; and also modify the prior by restricting it onto a constant neighborhood around the estimate. Due to the flexibility in selecting the prior distribution for MF-WGF, such a modification will have a minimal impact on the implementation. A second technical strategy is to further impose a dissipative condition (see Raginsky et al. (2017) for definition). A dissipative condition is commonly made to guarantee the long term stability of sampling algorithms such as Langevin dynamics Raginsky et al. (2017) as it causes most probability mass absorbed into a constant neighborhood of θ^* after a number of iterations; then the behavior of the algorithm inside this neighborhood is driven by the local convexity of the potential. Since Theorem 3 tells that there is at most $O\big(e^{-cn}\big)$ probability mass of $\widehat{q}_\theta$ outside this neighborhood, we expect

our current analysis to be valid up to an extra $O\big(e^{-cn}\big)$ remainder term. Due to the significant complexity of our current proof with global convexity, we will leave a systematic study on such an extension in a separate work.

Assumption B.2 (smoothness of population-level potential)**.** *There exists $L > 0$ such that U is L-smooth, which is defined by*

$$U\big((1-t)\theta + t\theta'\big) \geqslant (1-t)\,U(\theta) + t\,U(\theta') - \frac{L}{2}\,t(1-t)\|\theta - \theta'\|^2 \tag{31}$$

for all $t \in [0,1]$ and $\theta, \theta' \in \Theta$.

A smoothness condition on the objective function is usually necessary to prove exponential convergence of a coordinate descent-type optimization algorithms Wright (2015); Wright and Recht (2022). When U is twice differentiable, the above Assumption B.2 is equivalent to $\big\|\nabla^2 U(\theta)\big\|_{\mathrm{op}} \leqslant L$ for all $\theta \in \Theta$. In our proof, we can slightly relax this condition since we only require $\|\nabla_j U(\theta_j, \theta_{-j}) - \nabla_j U(\theta_j, \theta'_{-j})\| \leqslant L\|\theta_{-j} - \theta'_{-j}\|$, where ∇_j is the gradient with respect to the jth component of U. This inequality is equivalent to $\big\|\nabla_j \nabla_{-j} U(\theta)\big\|_{\mathrm{op}} \leqslant L$ when U is twice differentiable and is weaker than Assumption B.2.

To show that the sample-level potential U_n is uniformly close to its population version U and inherits the convexity and the smoothness of U (see the proof of Theorem 5 in Appendix C.3), we need the following assumption which characterizes the continuity and the sub-exponential tail of the (higher-order) derivatives of the log-likelihood functions.

Assumption B.3 (regularity of log-likelihood function)**.** *The log-likelihood function $\log p(x \mid \theta)$ is twice differentiable with respect to $\theta \in \Theta$. Let X denote a sample generated from the true distribution $p(\cdot \mid \theta^*)$. Then, the following regularity assumptions hold.*

1. *The Lipschitz constant (relative to the matrix operator norm) of the log-likelihood Hessian*

$$J(X) := \sup_{\theta \neq \theta'} \frac{\big\|\nabla^2 \log p(X \mid \theta) - \nabla^2 \log p(X \mid \theta')\big\|_{\mathrm{op}}}{\|\theta - \theta'\|}$$

   satisfies $\mathbb{E}_{\theta^*}[J(X)] < J_*$ for some finite J_*.

2. *For any $v \in B_{\mathbb{R}^d}(0,1)$ and $\theta \in \Theta$, $\big\langle v, \nabla^2 \log p(X \mid \theta) v \big\rangle$ is sub-exponential with parameter σ_5. In particular, a sufficient condition for this to be true is $\big\|\nabla^2 \log p(X \mid \theta)\big\|_{\mathrm{op}}$ being sub-exponential for any $\theta \in \Theta$.*

The first part of this assumption can be checked by controlling the third order derivatives of $\log p(X \mid \theta)$ with respect to θ when it is sufficiently smooth; the second part can be verified by directly calculating the first order Orlicz norm of $v^T \nabla^2 \log p(X \mid \theta)\, v$, and can also be extended to a bounded ψ_α (Orlicz) norm for some $\alpha > 0$.

Theorem 5 (MF-WGF without latent variables)**.** *Suppose Assumptions B.1–B.3 hold, and $\log \pi_\theta$ is twice differentiable. Then there is a universal constant $C > 0$, such that for any fixed $\eta \in (0,1)$, the following inequality holds with probability at least $1 - \eta$,*

$$W_2^2(q^{(k)}, \widehat{q}) \leqslant \big(1 + 2\tau\lambda_{lb} - L_{ub}^2 \tau^2 m\big)^{-k} W_2^2(q^{(0)}, \widehat{q}), \quad k \geqslant 1,$$

when $n \geqslant Cd\log d \cdot \max\big\{\log J_*/\log d, \log(R\sigma_5/\eta), 1\big\}$ *and the step size* τ *satisfies*

$$\tau^{-1} \geqslant \sqrt{m}L_{ub}, \quad \textit{and} \quad 1 + 2\tau\lambda_{lb} - L_{ub}^2\tau^2 m \geqslant 0,$$

where

$$\lambda_{lb} \coloneqq n\lambda - \lambda_{\max}(\nabla^2 \log \pi_\theta) - \sigma_5^2\sqrt{\frac{Cd\log n}{n} \cdot \max\Big\{\frac{\log J_*}{\log d}, \log\frac{R\sigma_5}{\eta}, 1\Big\}},$$

$$L_{ub} \coloneqq nL - \lambda_{\min}(\nabla^2 \log \pi_\theta) + \sigma_5^2\sqrt{\frac{Cd\log n}{n} \cdot \max\Big\{\frac{\log J_*}{\log d}, \log\frac{R\sigma_5}{\eta}, 1\Big\}},$$

and $\lambda_{\max}(\nabla^2 \log \pi_\theta)$ *and* $\lambda_{\min}(\nabla^2 \log \pi_\theta)$ *are the largest and the smallest eigenvalues of the Hessian matrix* $\nabla^2 \log \pi_\theta$ *in* Θ *respectively. In particular, if we take* $\tau = \lambda_{lb}/(L_{ub}^2 m)$*, then*

$$W_2^2(q^{(k)}, \widehat{q}) \leqslant \Big(1 + \frac{\lambda_{lb}^2}{L_{ub}^2 m}\Big)^{-k} W_2^2(q^{(0)}, \widehat{q}), \quad k \geqslant 1. \tag{32}$$

The proof of this theorem does not require utilizing the concentration property on $\widehat{q}_\theta$ as stated in Theorem 3, and the exponential convergence is solely driven by the convexity of population level potential U. However, when an effective potential U_n varies across iterations, which is the case in MF-WGF for Bayesian latent variable models, the concentration property becomes essential to manage the fluctuation of U_n.

Equation (32) implies that $O\big(\frac{mL_{ub}^2}{\lambda_{lb}^2}\log\big(\frac{1}{\varepsilon}\big)\big) = O\big(\frac{mL^2}{\lambda^2}\log\big(\frac{1}{\varepsilon}\big)\big)$ iterations are sufficient for the algorithm to achieve an accuracy of $\varepsilon \in (0,1)$ in computing $\widehat{q}$. This iteration complexity matches a typical iteration complexity of coordinate gradient descent for minimizing a strongly convex and smooth function in the Euclidean space when taking the step size in the order of $O\big(\frac{\lambda}{L^2 m}\big)$ as we do (see, e.g., Theorem 6.3 of Wright and Recht (2022)). However, analyzing the coordinate proximal gradient descent in a product Wasserstein space presents some unique challenges.

In the Euclidean space, $\|(x^{(k+1)} - x^{(k)}) - (\widehat{x} - x^{(k)})\| = \|x^{(k+1)} - \widehat{x}\|$ holds for any arbitrary points $\widehat{x}$, $x^{(k)}$, and $x^{(k+1)}$. However, we only have $\big\|T_{q_j^{(k)}}^{q_j^{(k+1)}} - T_{q_j^{(k)}}^{\widehat{q}_j}\big\|_{L^2(q_j^{(k)};\Theta_j)}^2 \geqslant W_2^2(q_j^{(k+1)}, \widehat{q}_j)$ due to the positive curvature of the Wasserstein space. This difference indicates that we have to evaluate the change of KL divergence along the generalized geodesics (see Appendix A.4 for a definition) connecting $q_j^{(k+1)}$ and $\widehat{q}_j$ with the base measure $q_j^{(k)}$, rather than the geodesics connecting $q_j^{(k+1)}$ and $\widehat{q}_j$.

In the proof, the convexity of the potential function $U(\theta)$ plays two roles: (1) allowing us to apply Lemma A.4 to control the subdifferential; (2) deriving a quadratic growth property of the KL-divergence functional, i.e.

$$D_{\mathrm{KL}}(q_1 \otimes \cdots \otimes q_m \,\|\, \pi_n) - D_{\mathrm{KL}}(\widehat{q}_1 \otimes \cdots \otimes \widehat{q}_m \,\|\, \pi_n) \gtrsim W_2^2(q, \widehat{q}_\theta), \quad \forall\, q_j \in \mathscr{P}_2^r(\Theta_j),\ j \in [m].$$

L-smoothness of $U(\theta)$ helps guarantee that the KL divergence between $q^{(k)}$ and π_n is decreasing when the step size is small. The detailed proof is postponed to Appendix C.3.

Bayesian latent variable models. Recall that with the presence of latent variables, the MF-WGF algorithm can be summarized by the following iterative updating rule: for $k = 0, 1, \ldots,$

$$q_\theta^{(k+1)} = \operatorname*{argmin}_{q_\theta} \; V_n(q_\theta \,|\, q_\theta^{(k)}) + \frac{1}{2\tau} W_2^2(q_\theta,\, q_\theta^{(k)}),$$

where for any $q'_\theta \in \mathscr{P}(\theta)$, the (sample) energy (or KL divergence) functional $V_n(\cdot \,|\, q'_\theta)$ is defined as

$$V_n(q_\theta \,|\, q'_\theta) := n\, \mathbb{E}_{q_\theta}\big[U_n(\theta;\, q'_\theta)\big] + D_{\mathrm{KL}}(q_\theta \,||\, \pi_\theta),$$

and $U_n(\cdot\,;\, q'_\theta)$ is the (sample) potential function given in (26). The corresponding population version of the potential is

$$U(\theta;\, q'_\theta) = -\int_{\mathbb{R}^d} \Big\{ \sum_{z=1}^{K} \log p(x, z \,|\, \theta)\, \Phi(q'_\theta, x)(z) \Big\}\, p(\mathrm{d}x \,|\, \theta^*). \tag{33}$$

The main difficulty in analyzing this MF-WGF algorithm is that the energy functional $V_n(\,\cdot \,|\, q_\theta^{(k)})$ determining $q_\theta^{(k+1)}$ also depends on the previous iterate $q_\theta^{(k)}$. With a time-independent energy functional, whose global minimizer denoted as π^*, we may directly apply Theorem 1 with $\pi = \pi^*$ to prove the contraction of the one-step discrete WGF towards π^*. However, by directly applying Theorem 1 with π therein being the minimizer of $V_n(\,\cdot \,|\, q_\theta^{(k)})$, we can only prove the one-step contraction of MF-WGF towards this minimizer, which changes over iteration count k and is generally different from the target $\widehat{q}_\theta$.

To overcome this difficulty in the convergence analysis, we may introduce an accompanied population-level MF-WGF, defined as

$$\widetilde{q}_\theta^{(k+1)} = \operatorname*{argmin}_{q_\theta} \; V(q_\theta \,|\, \delta_{\theta *}) + \frac{1}{2\tau} W_2^2(q_\theta,\, \widetilde{q}_\theta^{(k)}), \tag{34}$$

obtained by replacing $q_\theta^{(k)}$ in $V_n(\,\cdot \,|\, q_\theta^{(k)})$ by the point mass measure $\delta_{\theta *}$ at θ^*, and the sample energy functional by its population counterpart

$$V(q_\theta \,|\, q'_\theta) := n\, \mathbb{E}_{q_\theta}\big[U(\theta;\, q'_\theta)\big] + D_{\mathrm{KL}}(q_\theta \,||\, \pi_\theta). \tag{35}$$

Since the energy function $V(\,\cdot \,|\, \delta_{\theta *})$ in the population-level MF-WGF is time-independent, we can apply Theorem 1 or Corollary 2 to prove its contraction. Moreover, since $\widehat{q}_\theta$ is expected to be concentrated around θ^*, we may expect the trajectory of the sample-level MF-WGF to be close to that of the population-level MF-WGF under the same initialization.

Based on this discussion, a natural strategy to prove the convergence of the MF-WGF algorithm can be divided into two main steps: 1. control the difference between the sample-level iterates $\big\{q_\theta^{(k)} : k \geqslant 0\big\}$ and the population-level iterates $\big\{\widetilde{q}_\theta^{(k)} : k \geqslant 0\big\}$; 2. analyze the convergence of population-level MF-WGF (34). Our actual proof is slightly different from the above heuristics. Specifically, to simplify the proof, we do not explicitly bound the difference between the sample-level and population-level iterates, but use some population level quantities, such as potential U and energy functional V, to substitute their sample versions and properly control the resulting extra error terms in analyzing the sample iterates

(see Appendix C.4 for further details). In particular, the freedom of choosing an arbitrary π in Theorem 1 allows us to directly apply the theorem to analyze the sample-level MF-WGF by taking $\pi = \widehat{q}_\theta$; however, some careful perturbation analysis will be required for the proof to go through.

The following assumptions are needed to formally prove the convergence of MF-WGF.

Assumption C.1 (strong convexity of population-level potential)**.** *There exists some constant $r > 0$, such that for any $\mu \in B_{\mathbb{W}_2}(\delta_{\theta*}, r) := \{\mu \in \mathscr{P}_2^r(\mathbb{R}^d) : W_2(\mu, \delta_{\theta*}) \leqslant r\}$, function $U(\cdot\,;\mu) : \Theta \to \mathbb{R}$ is λ-strongly convex, i.e.*

$$U\big((1-t)\,\theta + t\,\theta'\,;\,\mu\big) \leqslant (1-t)\,U(\theta,\mu) + t\,U(\theta',\mu) - \frac{\lambda}{2}\,t(1-t)\,\|\theta-\theta'\|^2 \tag{36}$$

for all $t \in [0,1]$ and $\theta, \theta' \in \Theta$. Moreover, parameter space $\Theta \subset \mathbb{R}^d$ is convex and contained a ball centered at the origin with radius R.

By the definition of $U(\theta;\mu)$ as defined in (33), a sufficient condition of Assumption C.1 is the uniform strong log-concavity of $p(x, z\,|\,\theta)$ for all $x \in \mathbb{R}^d$ and $z \in [K]$. Notice that this assumption only requires $U(\cdot\,;\mu)$ to be strongly convex when μ is close to $\delta_{\theta*}$, another approach to verify this assumption is to prove the strong convexity of $U(\cdot\,;\delta_{\theta*})$, combining with a uniform control $|U(\theta\,;\delta_{\theta*}) - U(\theta\,;\mu)|$ with $W_2(\delta_{\theta*},\mu) \leqslant r$ for all $\theta \in \Theta$. Then, when the radius $r > 0$ is chosen small enough, $U(\cdot\,;\mu)$ presents strong convexity uniformly for all $\mu \in B_{\mathbb{W}_2}(\delta_{\theta*}, r)$. We will verify this assumption for the two applications considered in Section 6. When the initialization $q_\theta^{(0)}$ is close enough to the $\delta_{\theta*}$, it can be proved by induction that any later iterates $q_\theta^{(k)}$ will stay in the same W_2 neighborhood. Therefore, we do not need $U(\cdot\,;\mu)$ to be strongly convex for all $\mu \in \mathscr{P}_2^r(\Theta)$. This assumption plays a similar role as Assumption B.1 for models without latent variable. Similarly, we expect that the strong convexity of $U(\theta\,;\mu)$ with respect to parameter θ can be relaxed to a local strong convexity within a neighborhood of θ^* with sufficiently small constant radius for all $\mu \in B_{\mathbb{W}_2}(0, r)$. Here, we simply assume the global strictly convexity of U in the current analysis to avoid these technicalities without affecting the convey of our main proof ideas.

To show that the sample-level potential U_n is uniformly close to its population version U and inherits the convexity property of U (see Lemma C.6 in Appendix C.4), we need the following assumption characterizing continuity and sub-Gaussianity of the (higher-order) derivatives of the log-likelihood functions with the latent variable and the observed data.

Assumption C.2 (regularity of log-likelihood function)**.** *The log-conditional-likelihood function $\log p(z\,|\,x,\theta)$ of the latent variable Z and the log-marginal-likelihood function $\log p(x\,|\,\theta)$ of the observation X are twice differentiable with respect to θ for all $z \in [K]$ and $x \in \mathbb{R}^d$. Let X denote a sample from the true underlying data generating distribution $p(\,\cdot\,|\,\theta^*)$, then the following properties hold.*

1. *For $i = 1, 2$, the random variable $S_i(X) := \sum_{k=1}^K \big\|\nabla \log p(k\,|\,X,\theta^*)\big\|_2^i$ has finite expectation; and $S_2(X)$ is sub-exponential with parameter $\sigma_3 < \infty$, i.e. $\mathbb{E}\exp\{\sigma_3^{-1}|S_2(X)|\} \leqslant 2$.*

2. *If we denote the Lipschitz constant of the log-likelihood Hessian by*

$$J_k(X) := \sup_{\theta\neq\theta'} \frac{\big\|\nabla^2 \log p(X,k\,|\,\theta) - \nabla^2 \log p(X,k\,|\,\theta')\big\|_{\mathrm{op}}}{\|\theta-\theta'\|},$$

then there exist some finite constant J_* such that $\sum_{k=1}^K \mathbb{E}_{\theta^*}[J_k(X)] \leqslant J_*$.

3. *For any $v \in B_{\mathbb{R}^d}(0,1)$ and $\theta \in \Theta$, $\sum_{k=1}^K p(k \,|\, X, \theta^*) \cdot \langle v, \nabla^2 \log p(X, k \,|\, \theta)\, v \rangle$ is sub-exponential with parameter σ_1. In particular, a sufficient condition for this to hold is $\left\| \nabla^2 \log p(X, k \,|\, \theta) \right\|_{\mathrm{op}}$ being sub-exponential for any $\theta \in \Theta$ and $k \in [K]$.*

4. *$\lambda(X) := \sup_{\theta \in \Theta, k \in [K]} \left\| \nabla^2 \log p(k \,|\, X, \theta) \right\|_{\mathrm{op}}$ is sub-exponential with parameter σ_2.*

Part 2 of Assumption C.2 can be checked by controlling the third order derivatives of $\log p(X, k \,|\, \theta)$ with respect to θ when it is sufficiently smooth; part 3 can be verified by directly calculating the ψ_1-norm of $\sum_k p(k \,|\, X, \theta^*)\, v^T \nabla^2_\theta \log p(X, k \,|\, \theta)\, v$. Parts 3 and 4 of the assumption can also be extended to a bounded ψ_α (Orlicz) norm for some $\alpha > 0$. Now we present our main theoretical result on the convergence of MF-WGF for Bayesian latent variable models.

Theorem 6 (MFVI with latent variables)**.** *Suppose Assumptions A.1–A.3 and C.1–C.2 hold, and $\log \pi_\theta$ is twice differentiable. Let γ denote the operator norm of the missing data Fisher information matrix $I_S(\theta^*)$, i.e. $\gamma = \left\| I_S(\theta^*) \right\|_{\mathrm{op}}$ where*

$$I_S(\theta^*) = \int_{\mathbb{R}^d} \sum_{z=1}^K p(z \,|\, x, \theta^*) \big[\nabla \log p(z \,|\, x, \theta^*) \big] \big[\nabla \log p(z \,|\, x, \theta^*) \big]^T p(x \,|\, \theta^*)\, \mathrm{d}x,$$

and recall that r is the radius of the W_2-ball in Assumption C.1. Assume $\kappa := \frac{\lambda}{\gamma} > 2$, and define

$$R_W := \min \left\{ \sqrt{\frac{\lambda(\kappa-2)}{32A(\kappa+3)}}, \frac{\lambda(\kappa-2)}{16C(\kappa+3)}, \frac{\lambda(\kappa-2)}{8B(\kappa+3)}, \frac{r}{3} \right\},$$

where explicit expressions of the constants A, B and C are provided in the proof. If the initial distribution $q_\theta^{(0)}$ satisfies

$$W_2(q_\theta^{(0)},\, \delta_{\theta^*}) = \sqrt{\mathbb{E}_{q_\theta^{(0)}} \big[\|\theta - \theta^*\|^2 \big]} \leqslant R_W,$$

and the sample size n is large enough (explicit lower bound of n provided in the proof), then the k-th iterate $q_\theta^{(k)}$ satisfies

$$W_2^2(\, q_\theta^{(k)}, \widehat{q}_\theta) \leqslant \left(1 - \frac{(\kappa-2)(3\kappa+2) - \frac{2(3\kappa+2)}{n\gamma} \lambda_{\max}(\nabla^2 \log \pi_\theta)}{(4\kappa^2+\kappa-2) - \frac{2(3\kappa+2)}{n\gamma} \lambda_{\max}(\nabla^2 \log \pi_\theta) + \frac{2(3\kappa+2)}{n\tau\gamma}} \right)^k W_2^2(q_\theta^{(0)}, \widehat{q}_\theta).$$

with probability at least $1 - \frac{2}{\log n} - ne^{-\frac{\sqrt{n}\lambda(\kappa-2)}{4d\sigma_1(3\kappa+2)}} - 2e^{3d - \frac{cn\sigma_3\gamma(\kappa-2)}{4(3\kappa+2)}} - 2e^{-cn^{1/6}\sigma_2^{-1}} - 4e^{-cn^{1/6}\sigma_3^{-1}}$ for some universal constant $c > 0$. Again, $\lambda_{\max}(\nabla^2 \log \pi_\theta)$ is the largest eigenvalue of the Hessian matrix $\nabla^2 \log \pi_\theta$ in Θ. This means MF-WGF algorithm has exponential convergence towards the MF approximation $\widehat{q}_\theta$.

Remark 7. *As discussed previously, the proposed MF-WGF algorithm can be treated as a distributional version of the classical EM algorithm, which is known to converge exponentially fast only under certain conditions, such as a well-conditioned Fisher Information matrix (Dempster et al., 1977). Otherwise, EM enters a singular regime with slower, polynomial convergence (Dwivedi et al., 2020), such as in low signal-to-noise (SNR) settings (Dwivedi et al., 2020). The resulting parameter estimation convergence rate also becomes slower than the root-n parametric rate. Since a lower bound on the SNR is often required for exponential rates (Balakrishnan et al., 2017), we adopt similar assumptions to ensure faster convergence of MF-WGF.*

The contraction factor provided in the theorem decreases as the step size τ increases, with limit $1 - \frac{(\kappa-2)(3\kappa+2)}{4\kappa^2+\kappa-2}$ as $\tau \to \infty$. As we argued in Section H.1, JKO-scheme can be viewed as an implicit scheme in Wasserstein space. In the Euclidean setting, an implicit Euler scheme converges without any restriction on the step size. Similarly, we do not need any restriction on τ in our theory, and the existence of the solution of the optimization problem (19) for any $\tau > 0$ is proved in (Proposition 8.5, Santambrogio, 2015).

However, in practice, we need the k-th step size to satisfy $\tau_k \leqslant \frac{1}{2L_k+1}$ to guarantee the convergence of discretized Langevin SDE scheme. Here L_k denotes the Lipschitz constant of $n\nabla U_n(\cdot\,; q_\theta^{(k-1)}) - \nabla \log \pi_\theta$ (see Lemma H.1 and its proof for more details). The extra factor n is due to the leading multiplicative factor n in the definition of V_n. As a consequence, the theoretical upper bound requirement of step size is of order $O(n^{-1}\|\nabla^2 U_n(\cdot\,; q_\theta^{(k-1)})\|_{\mathrm{op}}^{-1})$, which matches the typical requirement of step sizes in gradient descents for empirical risk minimization.

The EM algorithm can be seen as a specific instance of our approach when the distributions in the MF family are further restricted to point mass measures. Thus, it is not surprising that our algorithm can only guarantee local convergence as the EM algorithm. Here, our Assumption C.1 does not directly impose local convexity on the population version of the negative log-likelihood. Instead, we focus on the population version of a distributional counterpart of the standard Q-function in the EM algorithm, which is defined in (33). In other words, our assumption allows the negative log-conditional likelihood function of observed data x given each latent variable value z to be non-convex in θ, as long as their weighted average remains convex; this assumption also does not require the conditional posterior of the parameter θ given latent variables Z^n to be log-concave. A similar local convexity assumption is made in the recent refined analysis of the EM algorithm by Balakrishnan et al. (2017). In addition, although our algorithm requires initializing in a neighborhood of the solution, the neighborhood radius from our theory is a *constant*, independent of sample size n, as opposed to a radius decreasing in n. This means that any initial estimator that is consistent can lead to a good initialization for our algorithm. While it may be feasible to relax this local convexity assumption, the primary focus of this work is not to enhance the existing convergence analysis of the EM algorithm but to show that many of the desirable properties associated with point estimators in frequentist literature can be extended to Bayesian cases. A primary message we wish to convey is that the computational framework of the Wasserstein gradient flow aligns well with existing analyses in conventional optimization (over Euclidean space) literature. This framework can leverage the inherent "convexity" structure to guarantee the convergence of certain MF algorithms.

In contrast, it is unclear whether the traditional CAVI algorithm for MF implementation (despite its limitation of requiring conditional conjugacy) can benefit from convexity, given that it can be interpreted as a gradient flow with respect to the KL divergence rather than the Wasserstein metric.

Our result requires an informative initialization which exists as long as $\kappa > 2$. This lack of global convergence is due to the following two reasons: (1) the strong convexity (C.1) is only required to hold in a neighborhood of δ_{θ^*}; (2) MF-WGF is an EM-type algorithm—it is known that, even in the Euclidean case, the EM algorithm converges to the true parameter with high probability when the initialization is good enough, but may converge to bad local optima with an uninformative initialization Balakrishnan et al. (2017). In practice, in order to choose a $q_\theta^{(0)}$ to satisfy the initial condition, we can run a simple and fast algorithm to get a consistent estimator of the parameter in order to construct a good initialization before applying MF-WGF. For example, in clustering problems, we may apply the EM algorithm or the K-means method to derive pilot estimates of all parameters; then, we may add independent noises with constant order variance to the previous estimates to generate i.i.d. particles inducing an initialization $q_\theta^{(0)}$.

Our proof of the theorem is based on an induction argument, by repeatedly applying Theorem 1 to analyze the evolution of one-step discretized WGF (19) for minimizing the energy functional $V(q_\theta \,|\, q_\theta^{(k)})$ whose form changes over the iteration count k. When sample size n is sufficiently large, the prior tend to have diminishing impact on the algorithm. If $\lambda \gg \gamma$ is also satisfied, then the derived algorithmic contraction rate is roughly of order $\mathcal{O}(\gamma/\lambda)$. Interestingly, γ reflects the amount of missing data information (by viewing latent variables as missing data), since recall that γ is defined as the operator norm of the missing data Fisher information $I_S(\theta^*)$; while λ corresponds to the complete data information, since it provides a lower bound to the complete data Fisher information $I_C(\theta^*)$ as the Hessian matrix of potential $U(\,\cdot\,;\,\delta_{\theta^*})$ at the point mass measure at θ^*. In comparison, the algorithmic contraction rate of the classical EM algorithm has a local contraction rate bounded by the largest eigenvalue of $\big[I_C(\theta^*)\big]^{-1} I_S(\theta^*)$ Dempster et al. (1977); and is consistent with the derived contraction rate of our MF-WGF algorithm viewed as a distributional extension of the EM.

By drawing an analogue from the local contraction rate of the EM algorithm, we believe that by incorporating some local geometric structures into the algorithm and our theoretical analysis, the current technical assumption $\lambda > 2\gamma$ can also be weakened to $I_C(\theta^*) \succeq a\, I_S(\theta^*)$ for all $\mu \in B_{\mathbb{W}_2}(0, r)$ and any constant $a > 1$. For example, we may use a weighted Euclidean norm, defined through $\|x - y\|_I^2 = (x-y)^T \big[I_C(\theta^*)\big]^{-1}(x-y)$, to substitute the isotropic Euclidean norm $\|x - y\|$ when defining the W_2 distance (1) and the one-step minimization movement scheme (19). With this substitution, we may define the strongly convexity coefficient of $U(\cdot\,;\mu)$ to be with respect to the $\|\cdot\|_I$ metric in the theoretical analysis, so that the key matrix $\big[I_C(\theta^*)\big]^{-1} I_S(\theta^*)$ will naturally appear when analyzing the contraction of the discrete gradient flow using Theorem 1. We leave a formal methodological and theoretical investigation about this improvement as a future direction.

The block MF approximation to posteriors in Bayesian latent variable models becomes accurate when the dependence between the parameter θ and latent variables Z^n is weak; or more formally, when the missing data Fisher information matrix $I_S(\theta^*)$ is small, such

that the latent variable distributions are not sensitive to perturbations or changes in the parameter θ. In fact, it is proved in Han and Yang (2019) that under this block MF, the marginal variational distribution $\widehat{Q}_\theta$ of the parameter θ approaches $N\big(\theta^{\mathrm{MLE}}, (nI_C(\theta^*))^{-1}\big)$ as the sample size n approaches infinity, where θ^{MLE} denotes the maximum likelihood estimator of θ, $I_C(\theta^*) = I_S(\theta^*) + I(\theta^*)$ is the complete data Fisher information and $I(\theta^*)$ denotes the (marginal) Fisher information matrix. In comparison, the classical Bernstein von-Mises theorem shows that the exact marginal posterior distribution of θ is close to $N\big(\theta^{\mathrm{MLE}}, (nI(\theta^*))^{-1}\big)$. Therefore, the block MF provides a good approximation to the target posterior distribution if and only if $I_S(\theta^*)$ is small. As an interesting implication, our Theorem 6 on the convergence of MFVI also suggests that the computational efficiency of MFVI improves as the statistical difficulty of approximating the joint posterior via MFVI decreases.

5. Computation

Note that both updating formulas (16) and (19) require solving the JKO scheme (2) when specializing $\mathcal{F}$ to be the KL-divergence type functional $\mathcal{F}_{\mathrm{KL}}$, i.e.

$$\rho_{k+1}^\tau = \operatorname*{argmin}_{\rho \in \mathscr{P}_2^r} \underbrace{\int V \,\mathrm{d}\rho + \int \rho \log \rho}_{\mathcal{F}_{\mathrm{KL}}(\rho)} + \frac{1}{2\tau} W_2^2(\rho, \rho_k^\tau). \tag{37}$$

In this section, we will consider and compare two numerical methods for numerically solving (37): particle approximation via SDE/diffusion and function approximation (FA) approach based on neural networks.

SDE approach. Recall that the JKO scheme (37) for KL divergence is an implicit scheme for discretizing the Fokker–Planck equation (7), which is also known as the WGF of $\mathcal{F}_{\mathrm{KL}}$. According to Section 2.4, the WGF of $\mathcal{F}_{\mathrm{KL}}$ starting from ρ_0 is the evolution of the following Langevin stochastic differential equation,

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad X_0 \sim \rho_0. \tag{38}$$

This connection between SGD and WGF motivates one to discretize the WGF by discretizing its corresponding SDE, and approximate the solution ρ_{k+1}^τ of the JKO scheme (37) by the evolution of the discretized SDE through the empirical measure of particles which satisfy the following updating formula,

$$X_b^{(k+1)} - X_b^{(k)} = -\nabla V(X_b^{(k)})\tau + \sqrt{2\tau}\eta_b^{(k)}, \quad b \in [B] \tag{39}$$

where $\big\{X_b^{(k)} : b \in [B]\big\}$ are B samples generated from ρ_k^τ, and $\eta_b^{(k)}$ are i.i.d. samples generated from $\mathcal{N}(0, I)$. This recursive equation is the discretized representation of the SDE (38) for approximating (37), and ρ_{k+1}^τ can be approximated by the empirical distribution of $\big\{X_b^{(k+1)} : b \in [B]\big\}$. See Appendix H.1 for further discussion about particle approximation and a numerical error analysis of its implementation via SDE.

FA approach. The function approximation method converts the JKO scheme (37) into an optimization problem over the function space. Note that finding the solution ρ_{k+1}^τ of (37) is equivalent to finding a transport map T such that $T_\# \rho_k^\tau$ minimizes (37). To be precise, we present the following theorem.

Theorem 8 (JKO scheme via function approximation)**.** *If $\rho_k^\tau \in \mathscr{P}_2^r$, and*

$$T_k^\tau = \operatorname*{argmin}_T \int V \circ T \,\mathrm{d}\rho_k^\tau - \int \log|\det \nabla T| \,\mathrm{d}\rho_k^\tau + \frac{1}{2\tau}\int \|T - Id\|^2 \,\mathrm{d}\rho_k^\tau, \tag{40}$$

then $\rho_{k+1}^\tau := (T_k^\tau)_\# \rho_k^\tau$ minimizes (37)*.*

We want to highlight a key property that the optimization problem (40) is unconstrained, although the last term $\int \|T - \mathrm{Id}\|^2 \,\mathrm{d}\rho_k^\tau$ corresponds to $W_2^2(\rho, \rho_k^\tau) = \min_{T,\,\mathrm{st}\, T_\#\rho = \rho_\tau^k} \mathbb{E}_\rho\big[\|X - T(X)\|^2\big]$, and requires the optimal transport map T_k^τ from ρ_k^τ to ρ_{k+1}^τ to be the gradient of a convex function according to Brenier's Theorem Brenier (1991) (see Appendix A.1 for more details). Most existing methods in the literature for numerically solving the JKO scheme, such as Mokrov et al. (2021), require solving a constrained optimization problem by restricting $T = \nabla\phi$ to the gradient of a convex function ϕ, where the convexity is imposed by using an input-convex neural network (ICNN) Amos et al. (2017). However, although ICNN is known to provide universal approximation to convex functions Chen et al. (2018), it is not clear whether its gradient also provides universal approximation to the gradients of convex functions. Moreover, based on our empirical observations, the inclusion of the convexity constraint tends to make the optimization problem particularly difficult to solve due to numerous local minima, extremely slow convergence and high sensitivity to tuning. On the contrary, Theorem 8 shows that solving the unconstrained optimization problem is equivalent to solving the JKO scheme, and even restricting T to be a gradient vector field is not necessary. The intuition is that, if a solution $\widetilde{T}_k^\tau$ to problem (40) is not the optimal transport map T_k^τ from ρ_k^τ to ρ_{k+1}^τ, then changing T_k^τ to $\widetilde{T}_k^\tau$ in the objective function (40) will strictly decrease the last transport cost term while keeping the rest terms unchanged. This contradicts to the optimality of $\widetilde{T}_k^\tau$. A formal proof is deferred to Appendix C.5.

In practice, the optimization problem in Theorem 8 over the function space can be solved by using function approximation methods, for example, based on (deep) neural networks. If we use T_k^τ to denote the transport map computed in the k-th iteration and choose an initial distribution ρ_0^τ that is easy to sample from, then we can approximate the objective functional in (40) up to arbitrary accuracy by Monte Carlo approximation via sampling from $\rho_k^\tau = \big[T_{k-1}^\tau \circ T_{k-1}^\tau \circ \cdots \circ T_0^\tau\big]_\# \rho_0^\tau$. Concretely, suppose $\{X_b^{(k)} : b \in [B]\}$ are B samples drawn from ρ_k^τ using the transport maps. We can compute the optimal transport map T_k^τ from ρ_k^τ to ρ_{k+1}^τ by solving

$$T_k^\tau = \operatorname*{argmin}_T \frac{1}{B}\sum_{b=1}^{B}\Big[V \circ T(X_b^{(k)}) - \log\big|\det \nabla T(X_b^{(k)})\big| + \frac{1}{2\tau}\big\|X_b^{(k)} - T(X_b^{(k)})\big\|^2\Big].$$

FA versus SDE. We recommend FA over SDE due to two major deficiencies arising in the SDE approach.

First, the SDE approach corresponds to the forward scheme that necessitates an upper bound on the step size τ to avoid divergence. In comparison, the FA approach is an

implicit scheme that does not diverge for any τ, provided that the associated optimization problem (40) is solved effectively. A typical upper bound on τ for SDE is proportional to the inverse smoothness parameter $\sup_{\theta\in\Theta} \|\nabla U_n(\theta)\|_{\mathrm{Lip}}^{-1}$. This restricts the use of larger step sizes for problems involving a fluctuating sample potential U_n, which, in turn, necessitates more iterations for SDE to converge to a reasonably good estimate (see Figures 3, 4 and 5).

Second, the SDE approach might introduce a systematic error that remains undiminished even with more iterations and number of particles. Specifically, it is known in the literature (e.g., Cheng et al. (2018); Chewi et al. (2021)) that applying a time-discretized SDE plus particle approximation to numerically compute a gradient flow over the space of all distributions suffers from two source of errors. One is the space and/or time discretization error due to a finite step size τ and a finite number B of particles; and another is the long term bias due to the mismatch between the limiting distribution of the Markov chain induced by the time-discretized SDE and the limiting distribution of the continuous time SDE. As a consequence, to attain an accuracy of $\varepsilon \in (0, 1)$ in the W_2 distance, both space and time complexities are $\mathcal{O}(\varepsilon^{-2})$ up to logarithmic factors. The second error resulting from the limiting bias, can be mitigated by incorporating a Metropolis-Hastings correction step. This gives rise to the Metropolis-adjusted Langevin algorithm (MALA, see e.g., Roberts and Tweedie (1996)). However, MALA is computationally much more expensive and it is not clear whether such a correction would be beneficial in our context. In comparison, each step of the FA approach is unbiased, meaning that any fixed point of the FA iterative formula (37) precisely gives a critical point to the target functional $\mathcal{F}_{\mathrm{KL}}$. As a consequence, unlike the SDE approach, the numerical error from earlier iterations of the FA approach will not accumulate provided the dynamics converge exponentially, which is the case in our scenario. Indeed, we observe this numerical issue with the SDE in our numerical experiments (e.g., refer to Figures 3, 4 and 5 in Section 6), where the optimization error from the SDE approach initially reduces but becomes unimprovable with increased iterations, stemming from the accumulated error and the limiting bias. In comparison, the optimization error in the FA approach continues to decline exponentially and shows a steeper decline in the logarithmic scale, suggesting a smaller contraction factor.

Nevertheless, using FA to solve the JKO scheme may still cause bias due to the limitation of the expressive ability of the chosen function class. This bias can be reduced by increasing the size of the function class, such as by choosing deeper and wider neural networks. However, in practice, larger and wider neural networks can make training the networks more challenging. Therefore, it is crucial to balance the bias and the trainability by choosing networks with proper sizes. In the examples presented in Section 6, the numerical results show that a small size of neural network is enough for the FA approach to have better performance than the SDE approach.

6. Examples

In this section, we apply our theoretical results to three representative Bayesian models and discuss their consequences. We also conduct some numerical studies to compliment the theoretical predictions. For simplicity, we only focus on Bayesian models with Gaussian noise. Nonetheless, we would like to emphasize that the proposed MF-WGF algorithm can also be easily applied to conduct MFVI for Bayesian models with non-Gaussian noise (Cabral

et al., 2024; Ma et al., 2019). To ensure theoretical guarantees for algorithmic convergence, it suffices to verify the Assumptions A.1–A.3 and C.1–C.2, as demonstrated in the examples presented in this section.

6.1 Bayesian linear regression

We consider Bayesian linear regression models as a representative example for Bayesian models without latent variables and verify the assumptions in Theorem 5. We consider a random design case where n i.i.d. pairs (X_i, y_i) are sampled from

$$y_i = \theta^T X_i + \varepsilon_i, \qquad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma) \qquad \text{and} \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \beta^2).$$

In this example, we assume both the coefficient θ and the variance β^2 are unknown parameters with the prior distribution $\pi(\theta, \beta^2)$. More specifically, we have

$$y_i \,|\, X_i, \theta, \beta^2 \sim \mathcal{N}(\theta^T X_i, \beta^2), \quad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \quad \text{and} \quad (\theta, \beta^2) \sim \pi(\theta, \beta^2),$$

where the covariance matrix Σ is positive definite. In the traditional setting, due to computational tractability, the prior of $\alpha := \beta^{-2}$ is usually Gamma distribution and the conditional prior distribution of $\theta \,|\, \beta^{-2}$ is chosen as a normal distribution. Here, we directly choose a uniform prior for θ and β^2, but our method can be easily implemented for all prior distributions that are absolutely continuous with respect to the Lebesgue measure.

Corollary 9. *Let Θ_α and Θ_θ be the parameter spaces of α and θ respectively. Assume $0 < \alpha_{lb} < \alpha < \alpha_{ub}$ for all $\alpha \in \Theta_\alpha$, and*

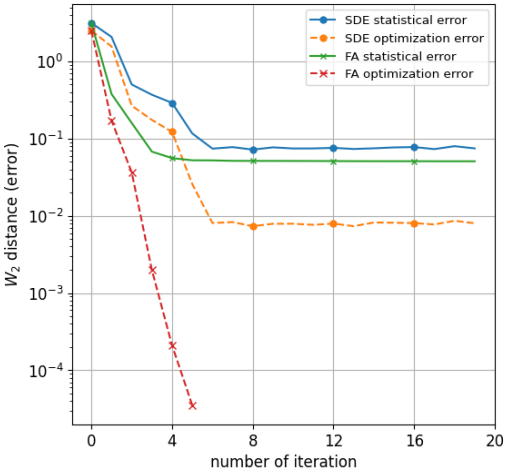
$$\sup_{\theta \in \Theta_\theta} \|\theta - \theta^*\| =: R_\theta < \sqrt{\frac{\lambda_1}{2\alpha_{ub}\lambda_d^2}}. \tag{41}$$

If $\lambda_1 I_d \preceq \Sigma \preceq \lambda_d I_d$, then we have

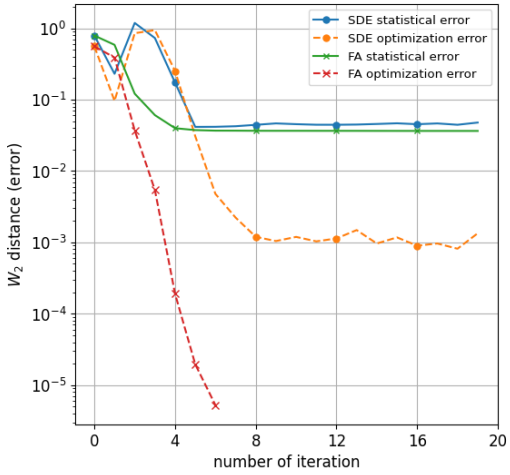
$$W_2^2(q_\theta^{(k)} \otimes q_\alpha^{(k)}, \widehat{q}_\theta \otimes \widehat{q}_\alpha) \leqslant \Big(1 + \frac{\lambda_{lb}^2}{L_{ub}^2 m}\Big)^{-k} W_2^2(q_\theta^{(0)} \otimes q_\alpha^{(0)}, \widehat{q}_\theta \otimes \widehat{q}_\alpha),$$

where

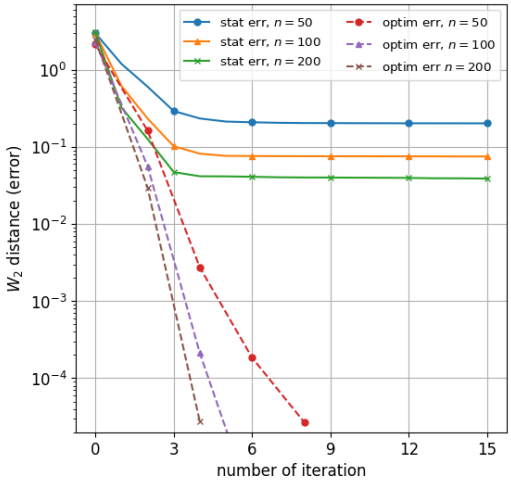
$$\begin{aligned}
\lambda_{lb} = {} & \frac{n(\frac{\lambda_1}{2\alpha_{ub}} - \lambda_d^2 R_\theta^2)}{\max\{\alpha_{ub}\lambda_1, \frac{1}{2\alpha_{lb}^2}\} + \lambda_d R_\theta} - \lambda_M(\nabla^2 \log \pi_\theta) \\
& - \sigma_5^2 \sqrt{\frac{Cd \log n}{n} \cdot \max\Big\{\frac{\log\big(2d + \frac{\alpha_{ub}}{\alpha_{lb}^4}\big)}{\log d}, \log \frac{R_\theta \sigma_5}{\eta}, 1\Big\}} \\
L_{ub} = {} & n\Big(\max\Big\{\alpha_{ub}\lambda_d, \frac{1}{2\alpha_{lb}^2}\Big\} + \lambda_d R_\theta\Big) - \lambda_m(\nabla^2 \log \pi_\theta) \\
& + \sigma_5^2 \sqrt{\frac{Cd \log n}{n} \cdot \max\Big\{\frac{\log\big(2d + \frac{\alpha_{ub}}{\alpha_{lb}^4}\big)}{\log d}, \log \frac{R_\theta \sigma_5}{\eta}, 1\Big\}}.
\end{aligned}$$



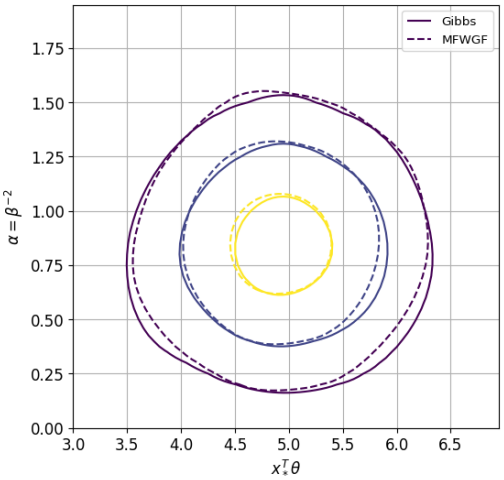
(a) Numerical error of θ.



(b) Numerical error of $\alpha = \beta^{-2}$.



(c) Numerical error for different sample size.



(d) Contour plot.

Figure 3: Numerical results in the Bayesian linear regression example with sample size $n = 100$, $\theta^* = (1, -2, 3)$, and $\beta^* = 1$. (a) and (b) Comparison of the numerical errors obtained by using the FA approach and the SDE approach of $\alpha = \beta^{-2}$ and θ. Both approaches have similar statistical errors, but different from the FA approach, the optimization error in the SDE approach converges to the approximation error after several iterations. A smaller step size leads to a smaller error when using the SDE approach; as a trade-off, it takes more iterations to converge. (c) Comparison of the numerical errors of θ obtained by using the FA approach with different sample sizes. When the sample size gets larger, the statistical error gets smaller, and the contraction rate does not change too much for sufficiently large sample size. (d) Comparison of contours from the joint posterior distribution of $(x_*^T\theta, \alpha)$ with $x_* = (-2, 1, 3)$ from Gibbs sampling versus their MF approximation output from MF-WGF. The MF approximation computed via MF-WGF is quite close to the true posterior.

Figure 3 summarizes the numerical results to support our theory in Bayesian models without latent variables. In this experiment, we consider the regression model with true parameters $\theta^* = (1, -2, 3)$ and $\beta^* = 1$. We choose the sample size $n = 100$ and use $B = 1000$ to approximate the posterior distribution. When applying the SDE approach, the particles which are used to approximate the distribution of α may go beyond the origin and become negative due to the unboundedness of Gaussian noise. To address this issue, we choose a threshold $\epsilon = 0.1$. At the end of each iteration, we add a projection step $\alpha_{b,\mathrm{proj}}^{(t)} = \alpha_b^{(t)} 1\{\alpha_b^{(t)} > \epsilon\} + \epsilon 1\{\alpha_b^{(t)} \leqslant \epsilon\}$ for all $b \in [B]$. We choose $\tau = 0.01$ for the SDE approach since it is the largest step size for SDE without incuring divergence, and use $\tau = 1$ for the FA approach.

Figure 3a presents the statistical error $W_2^2(\delta_{\theta^*}, q_\theta^{(k)})$ and the optimization error $W_2^2(\widehat{q}_\theta, q_\theta^{(k)})$ for the linear coefficient θ. Figure 3b shows the statistical error $W_2^2(\delta_{\alpha^*}, q_\alpha^{(k)})$ and the optimization error $W_2^2(\widehat{q}_\alpha, q_\alpha^{(k)})$ for the inverse of noise $\alpha = \beta^{-2}$. The increase of the statistical and optimization error in the first several iterates in the SDE approach is due to the existence of the projection step. As we can see, in both figures, the optimization error in the FA approach indicated by red dashed curves keeps decaying exponentially fast as predicted by our theory. However, in the SDE approach, the optimization error indicated by the orange dashed lines will finally be dominated by the approximation error and converge to quite large values compared with the FA approach. Choosing a smaller step size in the SDE approach can help decrease the approximation error. As a sacrifice, the algorithm takes more iterations to converge. In comparison, the statistical error indicated by solid curves has exponential decay at some initial period and then stabilizes in both approaches, which indicates the dominance of statistical error over optimization error in the later period.

Figure 3c studies the the effect of the sample size on the contraction rate and the statistical error when $\tau = 1$. In the plot, we can see that the statistical error indicated by the solid lines decreases when the sample size gets larger, which is consistent to the common knowledge in statistics. As for the contraction rate of the optimization error indicated by the dashed lines, increasing the sample size is helpful to get a smaller contraction rate when the sample size is small (compare $n = 50$ with $n = 100$); however, once a sufficient sample size has been acquired, further increase in sample size may result in little improvement (compare $n = 100$ with $n = 200$).

Figure 3d shows the contour plot of the joint posterior distribution of $(\alpha, x_*^T\theta)$ with $x_* = (-2, 1, 3)$, computed by Gibbs sampling, versus their MF approximation output by MF-WGF. From the plot, we see that the MF approximation is close to the true joint posterior distribution, meaning that prediction and its associated uncertainty quantification using MF tends to be accurate at x_*.

6.2 Repulsive Gaussian mixture model

In this example, we consider the Gaussian mixture model (GMM) as a simplest latent variable model that are widely used for clustering. We focus on the following (isotropic) Gaussian mixture model (GMM) with K components in $\mathbb{R}^d$,

$$p(x \,|\, m) = \sum_{k=1}^{K} w_i \mathcal{N}(x \,|\, m_i,\, \beta^2 I_d),$$

where the common covariance matrix is β^2 times the identity matrix I_d, $w = (w_1, \cdots, w_K) \in \mathbb{R}^K$ are the nonnegative mixing weight parameters satisfying $\sum_{k=1}^K w_k = 1$, and cluster centers $m = (m_1, \cdots, m_K) \in \mathbb{R}^{d\times K}$ are the primary parameters of interest. For theoretical convenience, we assume both nuisance parameters β^2 and w to be known; in our numerical results to follow, we treat w as unknown as well and use MF-WGF to approximate the joint posterior of (w, m). Our theory can also cover the model with unknown w and β^2 as long as the corresponding parameter spaces are convex, compact, and bounded away from zero.

As a common practice to simplify the likelihood computation and facilitate the interpretation, we introduce a latent variable $Z \in [K] := \{1, \cdots, K\}$ to indicate which underlying mixture component an observation X from GMM belongs to. Under this data augmentation, the full Bayesian latent variable model can be formulated as

$$\big[X_i \,\big|\, m,\, Z_i = k\big] \;\sim\; \mathcal{N}(m_k,\, \beta^2 I_d), \quad Z_i \overset{\text{i.i.d.}}{\sim} \text{Multi}([K], w), \quad \text{and} \quad m \;\sim\; \pi_m,$$

where π_m denotes the prior distribution over m. We apply a block MF approximation to the joint posterior distribution $\pi_n(m,\, Z^n)$ over parameter $m \in \mathbb{R}^{d\times K}$ and latent variables $Z^n = \{Z_1, \ldots, Z_n\}$, by using variational distributions of the form $q_{m,\, Z^n} = q_m \otimes q_{Z^n}$, to maximally preserve the dependence structure while retaining the computational tractability.

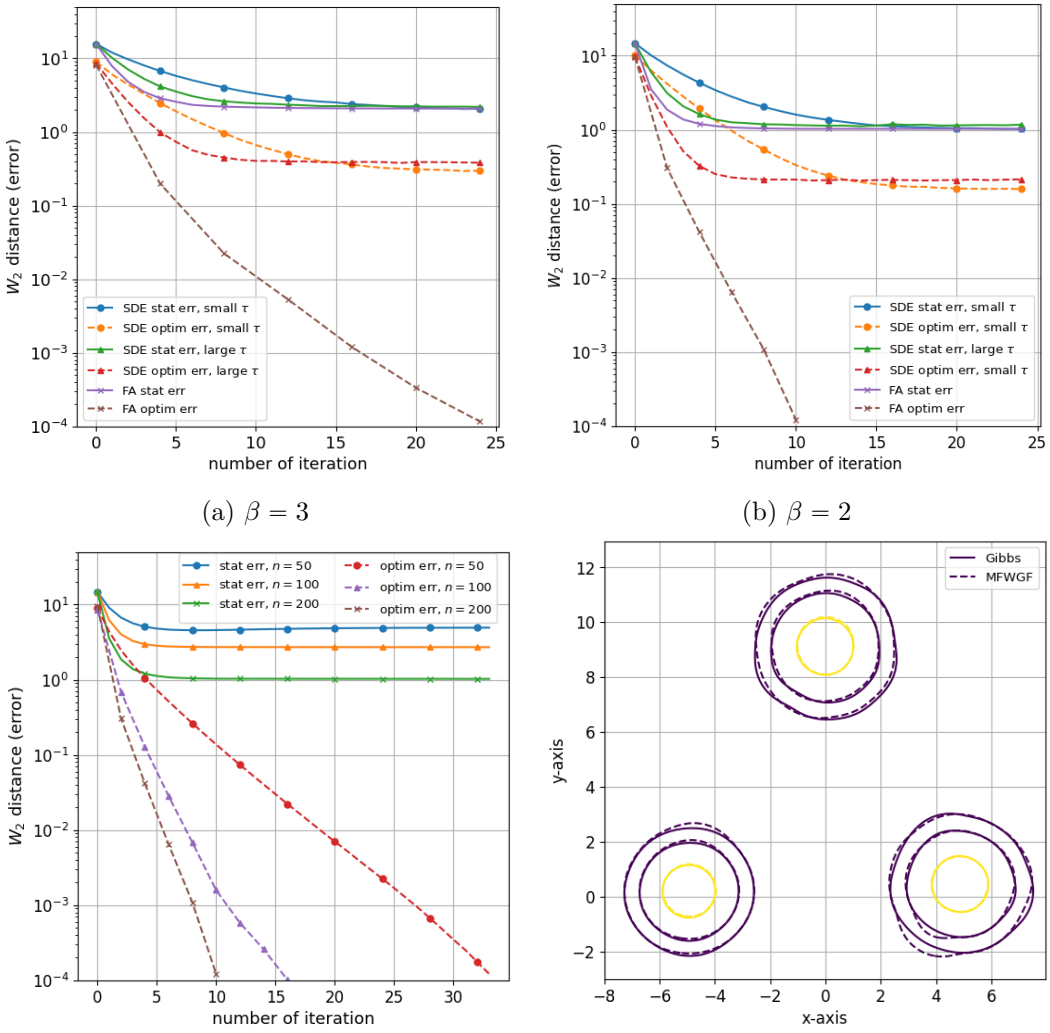
In the literature, there is a class of priors π_m, called repulsive priors Petralia et al. (2012); Xie and Xu (2020), that are preferable to use than independent priors over $\{m_k\}_{k=1}^K$. Let $d_{\min} = \min_{1\leqslant i<j\leqslant K} \|m_i - m_j\|$ denote the minimum distance between cluster centers. A typical repulsive prior takes the form of

$$\pi_m \propto g(m;\, g_0) \cdot \prod_{k=1}^K \mathcal{N}(m_k \,|\, 0, \sigma^2 I_d), \tag{42}$$

which modifies the independent priors with a repulsive function $g(m; g_0) = \frac{d_{\min}}{d_{\min}+g_0}$ that encourages the well-separatedness of cluster centers and reduces the potential redundancy of components. The complicated dependence structure introduced in the repulsive prior destroys the conditional conjugacy, making the standard coordinate ascent variational inference (CAVI, Bishop and Nasrabadi (2006)) algorithm for finding the best MF approximation $q_{m,\, Z^n}$ inapplicable. In comparison, the proposed MF-WGF can be easily applied in a straightforward manner. We want to emphasize that while we presented the repulsive prior as an illustrative example, our method is flexible and can be applied to various other priors without demanding additional restrictive conditions, such as the conditional conjugacy condition required by the widely-used CAVI algorithm. The following corollary proves the exponential convergence of MF-WGF and characterizes the explicit dependence of various problem characteristics on the contraction rate.

Corollary 10. *Let π_m be the prior of centers $m = (m_1, \cdots, m_K)$ satisfying Assumption A.2 and the parameter space $\Theta \subset B_{\mathbb{R}^{Kd}}(0, R)$. Assume the signal-to-noise ratio (SNR) $\kappa_{\text{SNR}} = \frac{d_{\min}}{\beta} > C$ for some constant $C = C(w, K) > 0$. Then, there exists constants $R_W = R_W(m^*, w, K, \beta)$ and $N = N(m^*, \beta, R_W, w, \pi_m, d)$ such that when $n > N$ and $W_2(\delta_{m^*}, q_m^{(0)}) \leqslant R_W$ (see (E.33) for explicit expressions), we have that*

$$W_2^2(\widehat{q}_m, q_m^{(k)}) \leqslant \rho^k\, W_2^2(\widehat{q}_m, q_m^{(0)}), \quad \forall\, k \in \mathbb{N}$$



(a) $\beta = 3$

(b) $\beta = 2$

(c) different sample size when $\beta = 2$

(d) Contour plot.

Figure 4: Numerical results of repulsive GMM with sample size $n = 200$ and repulsive parameter $g_0 = 1$. True centers are $m_1^* = (5, 0)$, $m_2^* = (0, 5\sqrt{3})$, and $m_3^* = (-5, 0)$ with weights $w^* = (0.27, 0.27, 0.46)$. (a) and (b): Comparison of the numerical errors obtained by using the FA approach and the SDE approach under different noise levels are shown in the figure. Statistical errors $W_2^2(q_\theta^{(t)}, \delta_{\theta*})$ of both approaches converge to similar values. However, the optimization error $W_2^2(q_\theta^{(t)}, \widehat{q}_\theta)$ in the SDE approach is dominated by the approximation error after several iterations. A smaller step size leads to a smaller error when using the SDE approach; as a trade-off, it takes more iterations to converge. (c) Comparison of the numerical errors obtained by using the FA approach with different sample sizes. Larger sample sizes yields smaller statistical error and faster convergence rate. (d) Comparison of contours from the marginal distributions of the cluster centers m_1, m_2 and m_3 from Gibbs sampling versus their MF approximation output from MF-WGF. The MF approximation computed via MF-WGF is quite close to the distribution computed via Gibbs sampling.

holds with probability at least $1 - \frac{3}{\log n}$. *Here, the contraction factor* ρ *takes the form of*

$$1 - \frac{(\zeta - 2)(3\zeta + 2)}{4\zeta^2 + \zeta - 2} \quad \textit{with} \quad \zeta = \frac{w_{\min}^2}{6K} \cdot \frac{\exp\{\kappa_{\mathrm{SNR}}^2/256\}}{2 + \kappa_{\mathrm{SNR}}^2},$$

as $n \to \infty$, *which monotonically decreases to* $\frac{1}{4}$ *as* $\kappa_{\mathrm{SNR}} \to \infty$.

Here, we want to make several remarks: 1. the repulsive prior (42) satisfies Assumption A.2, which then implies the exponential convergence of $q_m^{(k)}$ to $\widehat{q}_m$ in the W_2 metric (the proof can be found in Appendix E); 2. in practice, the compactness assumption on the parameter space is usually not necessary. Moreover, our algorithm can be straightforwardly extended to the setting where both cluster centers m and weights w are unknown as in the numerical studies shown below; 3. the lower bound of κ_{SNR} is not tight and can be improved. From our numerical experiments, a much smaller κ_{SNR} value is sufficient to ensure convergence to the global minimum. However, some lower bound on the κ_{SNR} is necessary to ensure the exponential convergence of the algorithm with a constant factor of contraction rate. With a low SNR, the model falls into the singular regime, and the EM algorithm (as well as our algorithm) may converge extremely slowly; see, for example, Dwivedi et al. (2020). This slow-convergence is natural since in the singular regime, the parameter itself becomes statistically non-identifiable (due to a near singular Fisher information matrix) and cannot be accurately estimated. The same remark also applies to the next mixture of regression example.

Figure 4 summarizes some numerical results to complement the theoretical predictions. In this experiment, we consider GMM with three classes centered at $m_1 = (5, 0)$, $m_2 = (0, 5\sqrt{3})$, and $m_3 = (-5, 0)$ with weights $w_1 = w_2 = 0.27$ and $w_3 = 0.46$. We choose the repulsive prior (42) with $g_0 = 1$ and $\sigma^2 = 10$. We let the sample size $n = 200$. For the SDE approach under both noise settings ($\beta = 2$ and $\beta = 3$), we choose the largest step size while trying not to increase the statistical error significantly. We construct the initialization by applying the K-means clustering to obtain an initial estimates of m. We use 1000 particles in the simulation in order to estimate the optimization (numerical) error $W_2^2(\widehat{q}_\theta, q_\theta^{(k)})$. Far less particles will be needed for conducting accurate inference on the model parameters. As indicated by Corollary 10, we define d/β as the signal to noise ratio (SNR) that characterizes the algorithmic convergence, and vary it in the simulation by tuning β. Since we are using the log-scale for the vertical axis, straight lines means our considered squared W_2 distance, either the optimization error $W_2^2(\widehat{q}_\theta, q_\theta^{(k)})$ or the statistical error $W_2^2(\delta_{\theta*}, q_\theta^{(k)})$, decays exponentially fast in the number of iterations, with the slope corresponding to the logarithm of the contraction rate.

In Figure 4a, we choose $\tau = 0.03$ and $\tau = 0.015$ for the SDE approach and $\tau = 0.08$ for the FA approach. Similar to the phenomenon in Figure 3, in the FA approach, the optimization error indicated by dashed curves keeps decaying exponentially fast as predicted by our theory; in the SDE approach, the optimization error will finally be dominated by the approximation error. Choosing a smaller step size can help decrease the approximation error but makes the algorithm take more iterations to converge. In comparison, the statistical error indicated by solid curves has exponential decay at some initial period and then stabilizes in both approaches, which indicating the dominance of statistical error over optimization error in the later period.

In Figure 4b, we choose $\tau = 0.025$ and $\tau = 0.01$ for the SDE approach and $\tau = 0.1$ for the FA approach. The same trend of curves as in Figure 4a is observed as well. Comparing with Figure 4a, we can see that a higher SNR (smaller β) corresponds to a faster decay, i.e. smaller contraction rate. SNR also encodes the statistical hardness of the problem: a higher SNR corresponds to a higher statistical error as the stabilized values of solid curves in the plots.

In Figure 4c, we consider the effect of the sample size on the contraction rate and the statistical error when $\tau = 0.1$. In the plot, we can see that the statistical error indicated by the solid lines decreases when the sample size gets larger, which is consistent to the common knowledge in statistics. As for the contraction rate of the optimization error indicated by the dashed lines, increasing the sample size is helpful to get a smaller contraction rate when the sample size is small (compare $n = 50$ with $n = 100$); however, once a sufficient sample size has been acquired, further increases in sample size may result in little improvement (compare $n = 100$ with $n = 200$).

Figure 4d shows the contour plots of the true posterior distribution of cluster centers (solid curves) versus their MF approximation output by MF-WGF (dashed curves), which are pretty close to each other.

6.3 Mixture of regression

We consider as the third illustrative example a finite mixture regression model (FMRM, Sung (2004); Viele and Tong (2002)), which can be viewed as an extension of the GMM by including covariates in the mixture formulation. In the standard (random-design) linear regression model, we observe n i.i.d. pairs $(y_i, X_i)_{i=1}^n$ from

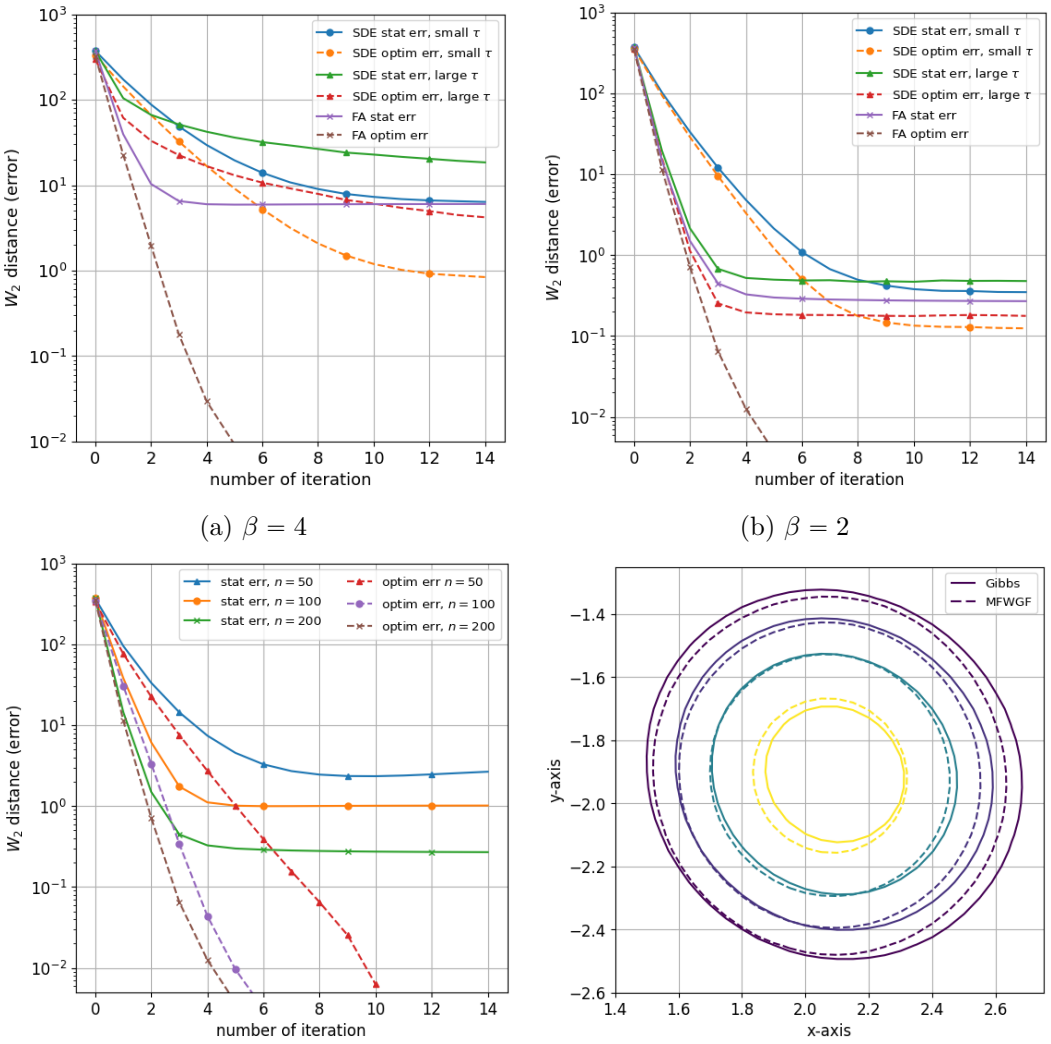
$$y_i = X_i^T \theta + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \beta^2)$$

where $X_i \in \mathbb{R}^d$ denotes the ith covariant vector, y_i is the ith response variable, $\theta \in \mathbb{R}^d$ is the unknown regression coefficient vector parameter of interest, and the Gaussian noise ε_i is independent of (X_i, Y_i). In our theoretical analysis, we assume X_i to be sampled from $\mathcal{N}(0, I_d)$. By introducing clustering structures on the conditional distribution of Y_i given X_i, we reach the FMRM. Concretely, we focus on the simple case with two equally weighted symmetric clusters, where each cluster is determined by its own regression coefficient vector in $\mathbb{R}^d$, as

$$\big[y_i \,\big|\, X_i, \theta, Z_i\big] \;\sim\; \mathcal{N}(Z_i X_i^T \theta, \beta^2), \quad Z_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\{1, -1\}, \quad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \quad \text{and} \quad \theta \;\sim\; \pi_\theta,$$

where π_θ denotes the prior of θ. In this model, we are interested in approximating the posterior distribution of θ. For the sake of parameter identifiability, we assume the first non-zero component of θ is positive (θ and $-\theta$ correspond to the same model). We also consider the block MF approximation by using $q_{\theta,\, Z^n} = q_\theta \otimes q_{Z^n}$ to approximate the joint posterior of $(\theta,\, Z^n)$, where $Z^n = \{Z_1, \ldots, Z_n\}$. The following corollary provides the algorithmic convergence of MF-WGF algorithm for computing the MF solution $\widehat{q}_\theta$.

Corollary 11. *Let π_θ be any prior satisfying Assumption A.2. If the SNR of the problem, defined as $\kappa_{\text{SNR}} = \frac{\|\theta^*\|}{\beta}$, is sufficiently large so that the ζ to be defined below satisfies $\zeta > 2$,*



(a) $\beta = 4$

(b) $\beta = 2$

(c) different sample size when $\beta = 2$

(d) Contour plot.

Figure 5: Numerical results in the mixture of regression example with sample size $n = 200$, and a 10-dimensional $\theta^* = (0, 1, 2, 3, 4, 0, -1, -2, -3, -4)$. (a) and (b): Comparison of the numerical errors obtained by using the FA approach and the SDE approach under different noise levels are shown in the figure. Both approaches have similar statistical errors $W_2^2(\widehat{q}_\theta, \delta_{\theta*})$, but different from the FA approach, the optimization error $W_2^2(q_\theta^{(t)}, \widehat{q}_\theta)$ in the SDE approach converges to the approximation error after several iterations. A smaller step size leads to a smaller error when using the SDE approach; as a trade-off, it takes more iterations to converge. (c) Comparison of the numerical errors obtained by using the FA approach with different sample sizes. When the sample size gets larger, the statistical error gets smaller. (d) Comparison of contours from the posterior distribution of the coefficient θ on the 3rd and the 8th coordinates from Gibbs sampling versus their MF approximation output from MF-WGF. The MF approximation computed via MF-WGF is quite close to the true posterior.

then there exists constant $N = N(\theta^*, \beta, d, \pi_\theta)$, such that as long as the initialization satisfies

$$W_2(\delta_{\theta^*}, q_\theta^{(0)}) \leqslant R_W = \frac{C'\beta^{-2}(\zeta-2)}{\Big[K^2\big(d\beta^{-3}(\kappa_{\mathrm{SNR}}^2+1)^{3/2}+1\big)+K^3\big(d\beta^{-6}(\kappa_{\mathrm{SNR}}^2+1)^3+1\big)\Big](\zeta+3)}$$

for some constants $C' > 0$, ζ and $n > N$, we have that

$$W_2^2(\widehat{q}_\theta, q_\theta^{(k)}) \leqslant \rho^k W_2^2(\widehat{q}_\theta, q_\theta^{(0)}) \quad \forall\, k \in \mathbb{N}$$

holds with probability at least $1 - \frac{3}{\log n}$. The contraction factor ρ takes the form of

$$1 - \frac{(\zeta-2)(3\zeta+2)}{4\zeta^2+\zeta-2} \quad \textit{with} \quad \zeta = \frac{(16+\kappa_{\mathrm{SNR}}^2)^{1/4}}{2174},$$

as $n \to \infty$, which is decreasing in κ_{SNR}.

The proof of the above corollary is postponed to Appendix E.3. Directly solving $\zeta > 2$ with ζ defined above provides a very loose bound of $\kappa_{\mathrm{SNR}} > 1.8 \times 10^7$. In fact, the lower bound requirement can be substantially improved to a positive constant less than 10 by numerically calculating an analytically intractable constant in our proof. More details are referred to the end of Appendix E.3.

Figure 5 shows the simulation results for implementing the mixture of regression model via MF-WGF. We set $\theta^* = (0, 1, 2, 3, 4, 0, -1, -2, -3, -4)$ in the data generative model and generate $n = 200$ i.i.d. samples. For $\beta = 4$, we choose $\tau = 0.2$ and $\tau = 0.05$ for the SDE approach and $\tau = 0.4$ for the FA approach. For $\beta = 2$, we choose $\tau = 0.02$ and $\tau = 0.01$ for the SDE approach and $\tau = 0.1$ for the FA approach. We vary the SNR value $\|\theta^*\|/\beta$ by changing the noise variance β^2. Similar to the GMM example, we observe nearly straight lines for the numerical error $\log W_2^2(\widehat{q}_\theta, q_\theta^{(k)})$ versus the iteration count, indicating the exponential convergence of the algorithm. Moreover, a higher SNR corresponds to a smaller contraction rate as predicted by our theory. The statistical error $W_2^2(\delta_{\theta^*}, q_\theta^{(k)})$, indicated by solid curves in the plot, is at first dominated by the optimization error, but afterwards dominates the latter and stabilizes. Figure 5a and Figure 5b compares the FA approach with the SDE approach and the numerical error under different noise levels. Figure 5c studies the affect of the sample size on the contraction rate and the statistical error when $\tau = 0.1$. Figure 5d shows the contour plots of the true posterior distribution of θ and its MF approximation.

7. Discussion

In this paper, we have proposed a general computational framework for realizing the mean-field variational approximation to Bayesian posteriors via Wasserstein gradient flows. We also applied the developed methods and theory to three concrete examples, linear regression model for Bayesian models without latent variables, and Gaussian mixture model and mixture of regression model for Bayesian latent variable models. Our analysis implies the exponential convergence of the algorithm given a good initialization.

We also expect the development of this paper can be extended to other varitional approximation schemes, and the theoretical results to hold under weaker assumptions. For instance, as we briefly remarked in Section 4.2, we may relax the global convexity condition on U relative to the parameter θ into a local one, and the condition on the "condition number" κ in Theorem 6 into a weaker one. It is also possible to use a pre-conditioned Wasserstein distance with cost function as a weighted Euclidean norm square in constructing the discrete-time Wasserstein gradient flow, so that the local geometric structure can be captured while updating the variational distribution. This variant can be viewed as the generalization of the usual quasi-Newton's method to the Wasserstein space, which may enjoy a faster rate of algorithmic convergence.

It is also of interest to investigate the accumulation of errors arising from Langevin Monte Carlo updates or neural network approximations when implementing the JKO scheme in the MF-WGF algorithm. One may possibly replace the JKO scheme with additional iterations of Langevin Monte Carlo and analyze the resulting error using tools from the sampling literature (e.g., (Chewi et al., 2024; Lee et al., 2021; Wibisono, 2018)). Alternatively, one may leverage results from neural approximation theory for shallow networks (e.g. (Proposition 10, Sreekumar and Goldfeld, 2022) or Fan et al. (2023)) to study the accumulation of neural network approximation errors.

Another potential topic is to extend our algorithm to the non-parametric setting. For example, we can try to extend the application of MF-WGF algorithm to Gaussian sequence model as studied in (Zhang and Gao, 2020), and Bayesian semiparametric models, such as partial linear models where the mean-field approximation is applied to bothnthe parameter of interest block and the nuisance parameter block (Bickel and Kleijn, 2012). We leave all these threads into future directions.

Acknowledgments

Yun Yang was partially supported by NSF grant DMS-2210717. The online supplement is available at `https://arxiv.org/pdf/2207.08074`.

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, 2020.

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

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

James R Anderson and Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. On the convergence of coordinate ascent variational inference. *arXiv preprint arXiv:2306.01122*, 2023.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

PJ Bickel and BJK Kleijn. The semiparametric bernstein-von mises theorem. *The Annals of Statistics*, pages 206–237, 2012.

Lucien Birgé. *Sur un théoreme de minimax et son application aux tests*. Univ. de Paris-Sud, Dép. de Mathématique, 1979.

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

David M Blei and Michael I Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

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

Rafael Cabral, David Bolin, and Håvard Rue. Fitting latent non-gaussian models using variational bayes and laplace approximations. *Journal of the American Statistical Association*, 119(548):2983–2995, 2024.

Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. *Foundations of Computational Mathematics*, 22(2):389–443, 2022.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal control via neural networks: A convex approach. *arXiv preprint arXiv:1805.11835*, 2018.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. 2001.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J Wainwright, Michael I Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of em. 2020.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023.

Charles W Fox and Stephen J Roberts. A tutorial on variational bayesian inference. *Artificial intelligence review*, 38(2):85–95, 2012.

Charlie Frogner and Tomaso Poggio. Approximate inference with wasserstein gradient flows. In *International Conference on Artificial Intelligence and Statistics*, pages 2581–2590. PMLR, 2020.

Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

Peter Hall, John T Ormerod, and Matt P Wand. Theory of gaussian variational approximation for a poisson mixed model. *Statistica Sinica*, pages 369–389, 2011a.

Peter Hall, Tung Pham, Matt P Wand, and Shen SJ Wang. Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, 39(5): 2502–2532, 2011b.

Wei Han and Yun Yang. Statistical inference in mean-field variational bayes. *arXiv preprint arXiv:1911.01525*, 2019.

Tommi S Jaakkola and Michael I Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, pages 283–294. PMLR, 1997.

Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

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.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

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

Kenneth Lange. *MM optimization algorithms*. SIAM, 2016.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Zhanyu Ma, Jalil Taghia, and Jun Guo. On the convergence of extended variational inference for non-gaussian statistical models. *arXiv preprint arXiv:1902.05068*, 2019.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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.

G. Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781. URL `https://books.google.com/books?id=IG7CGwAACAAJ`.

Manfred Opper and Ole Winther. A mean field algorithm for bayes learning in large feed-forward neural networks. *Advances in Neural Information Processing Systems*, pages 225–231, 1997.

John T Ormerod and Matt P Wand. Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21(1): 2–17, 2012.

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

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588. PMLR, 2018.

Francesca Petralia, Vinayak Rao, and David Dunson. Repulsive mixtures. *Advances in neural information processing systems*, 25, 2012.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pages 63–95. Springer, 1996.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. *arXiv preprint arXiv:2002.03035*, 2020.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226. PMLR, 2015.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Sreejith Sreekumar and Ziv Goldfeld. Neural estimation of statistical divergences. *Journal of machine learning research*, 23(126):1–75, 2022.

Hsi Guang Sung. *Gaussian mixture regression and classification*. Rice University, 2004.

DM Titterington and Bo Wang. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1 (3):625–650, 2006.

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

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *arXiv preprint arXiv:1209.4360*, 2012.

Ted Westling and Tyler H McCormick. Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. *arXiv preprint arXiv:1510.08151*, 1, 2015.

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.

David Wingate and Theophane Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.

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.

Fangzheng Xie and Yanxun Xu. Bayesian repulsive gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203, 2020.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, 48 (5):2575–2598, 2020.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, 2020.