On Representations of Mean-Field Variational Inference

Soumyadip Ghosh
IBM Research
Yorktown Heights, NY 10598
sghosh@us.ibm.com

Yingdong Lu
IBM Research
Yorktown Heights, NY 10598
yingdong@us.ibm.com

Tomasz Nowicki
IBM Research
Yorktown Heights, NY 10598
tnowicki@us.ibm.com

Edith Zhang
Applied Physics and Applied Mathematics
Columbia University, NY 10027
ejz2120@columbia.edu

Abstract

We obtain representations of Mean-Field Variational Inference (MFVI) in the forms of gradient flows on product spaces, quasilinear partial differential equations and McKean-Vlasov diffusion processes. These new interpretations not only provide new understanding of MFVI, but also allows us to conduct new analysis on its convergence.

1 Introduction

The connections between Bayesian inference and key concepts in analysis and probability such as gradient flows, partial differential equations (PDE) and diffusion processes have been established previously, [9] and [11]. As a consequence, Bayesian inference can be viewed as the minimizer to variational problems in multiple forms. Further, tools developed and extensively studied in the physics literature, for example minimization under statistical divergences, of Dirichlet energy etc., can be applied to Bayesian inference. These variational approaches not only provide deeper understanding but also more effective computational methods for the inference problem; for example, the connection to diffusion process can be explored to design Markov chain Monte Carlo algorithms for inference. Mean Field Variational Inference (MFVI) is a popular and powerful approximation technique for the Variational Inference (VI) problem. With a view towards bringing the same benefits, in this paper we describe how the solution of MFVI can also be represented as *a limit of a gradient flow*, *a solution to a homogeneous PDE* and *the stationary distribution of a diffusion process*.

The *Variational Bayes* (VB) [3] form of Bayes' rule expresses the posterior p as the minimizer of the Kullback-Leibler (KL) divergence D to itself, where $D(\xi\|\eta) := \mathbb{E}_\xi[\log(d\xi/d\eta)]$ for ξ and η,

$$p = \arg\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} D(\nu\|p) = \arg\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} \{\mathbb{E}_\nu[\log\nu] - \mathbb{E}_\nu[\log \mathsf{P}(\boldsymbol{x},\theta)]\} + \log Z. \tag{1}$$

Here, the set $\mathcal{P}(\mathbb{R}^d)$ contains absolutely continuous probability measures. Denote as $H(\nu) := -\mathbb{E}_\nu[\log\nu]$ the entropy of the measure ν, and $\Psi(\nu) := \mathbb{E}_\nu[-\log \mathsf{P}(\boldsymbol{x},\theta)]$ the expected negative log likelihood of the joint distribution $\mathsf{P}(\boldsymbol{x},\theta)$. The VB (1) minimizes the evidence lower bound [3] objective $J(\nu) := \Psi(\nu) - H(\nu)$. Equivalently, it maximizes $-J(\nu)$, balancing a high log likelihood $\Psi(\nu)$ under ν with a regularization term that desires a high entropy solution ν; see [11] for a comprehensive review of representations from other perspectives. MFVI algorithms seek among the joint distribution in a product form for over the d latent components $\theta = (\theta_1, \ldots, \theta_d)$ the one that is closest to the posterior distribution. As it turns out, this restriction poses some serious barriers

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

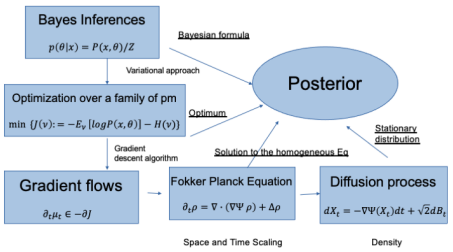

Figure 1: Representations of VB (1)

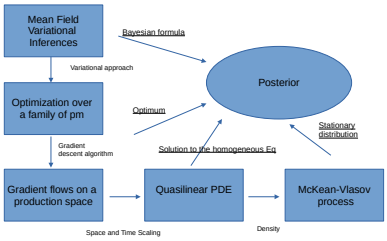

Figure 2: Representations of MFVI

for interpreting it as an analytical object in e.g. gradient flow and PDE. First, to properly define the right gradient flow, some classic definitions and concepts of gradient flow have to be extended to the product space, and not all of them are straightforward. With the gradient flow properly defined, we show that the MFVI converge to a gradient flow on product space as the step size goes to zero. Furthermore, this leads to a finite-time convergence rate result for MFVI under convexity assumptions, which is a first in the literature as far as we know. Second, the PDE that can reflect the dynamics of MFVI is a nonlinear second order parabolic equations, instead of the well-studied linear Fokker-Planck equation as in the Bayesian case. In the Bayesian case, many regularity (such as uniqueness) and invariant properties of the linear Fokker-Planck equation can be used to derive necessary results for gradient flow (see e.g. [9]) and diffusion processed. That luxury is not share in the case of MFVI. While the existence of the class of nonlinear equation is known to certain degree, the general uniqueness of the solution remains to be an open problem. To deal with this difficulty, we are able to secure contraction (thus uniqueness) for our gradient flow by utilizing the concept of geodesic convexity, and avoid the dependence of the uniqueness of the nonlinear PDE. Moreover, we show the uniqueness for a family of nonlinear PDE with the help of the gradient flow. This is a nice addition to the PDE literature, as a solid step towards an answer to the general uniqueness problem. Third, we identify the stochastic differential equation that characterizes the probabilistic evolution of MFVI, this supplies a foundation for new potential MCMC algorithms. We further demonstrate that it has terms explicitly depends on the law of the process, thus belongs to the family of McKean-Vlasov processes, which are actively studied in mean field game and mean field control, see,e.g. [6]. In addition to presenting the more detailed statement on each of the results below, we also illustrate them and their relationships in Figure 2, in comparison with the variational Bayesian inference counterpart in Figure 1.

Prior Work: Convergence analysis of the MFVI approximation is relatively less well established. [12] provide consistency results for MFVI procedures by establishing that point estimates of expectations of functions of the latent variables θ constructed using MFVI estimates of the posterior converge to the true value asymptotically as the size n of $\boldsymbol{x}$ grows under the assumption that the true latent variable takes a definite value. Recently, [10] presents a convergence analysis of MFVI where the components are further constrained to be Gaussian distributions and their mixtures, thus operating in the sub-manifold of $\mathcal{P}(\mathbb{R}^d)$ known as the Bures-Wasserstein manifold. [13] study MFVI for the special case where the posterior for $\theta = (\varphi, z_1, \ldots, z_n)$ is factored into a distribution with general support for φ and discrete distributions for z_i.

2 Representation of Bayesian Inference

Existence, uniqueness and convergence results for VB can be obtained from representations of (1) that exploit connections between Bayesian inference, differential equations and diffusion processes. [9] provided a seminal result that the gradient flow in Wasserstein space (the metric space $\mathcal{P}(\mathbb{R}^d)$ of probability measures endowed with 2-Wasserstein distance W_2, definitions below) of an objective function like (1) can be equivalently expressed as the solution to a Fokker-Planck (FPE) equation, which is a parabolic partial differential equation (PDE) on densities as L_1 functions. This key connection allow Bayes' rule to be expressed as minimum of various related functionals on different metric spaces: it can be viewed as the stationary solution of a gradient flow of J in the space W_2, as the stationary solution to an FPE in the L_1 space of density functions, and also corresponds to the stationary distribution of a diffusion process. Fig. 1 depicts these equivalent relationships.

3 MFVI Formulation

The mean field variational inference approximation of the posterior solves:

$$\min_{\nu \in \mathcal{Q}(\mathbb{R}^d) = \prod_{i=1}^d \mathcal{P}_i^2(\mathbb{R})} J(\nu), \tag{2}$$

with $J(\nu)$ treated as a functional on the product space of square-integrable measures $\mathcal{Q}(\mathbb{R}^d) = \prod_{i=1}^d \mathcal{P}_i^2(\mathbb{R})$. This is encoded via the constraint $\nu(\theta) = \prod_{i=1}^d \nu_i(\theta_i)$. With this constraint, $J(\nu)$ from (1) now takes the form $J(\nu) = -\int_{\mathbb{R}^d} \log \mathsf{P}(\boldsymbol{x}, \theta) \prod_{i=1}^d \nu_i(\theta_i) d\theta_i - \sum_{i=1}^d H(\nu_i)$ with $H(\nu_i) = -\int_{\mathbb{R}} \nu_i \log \nu_i d\theta_i$ is the entropy of the i-th component. To simplify notation, let $\nu_{-i} := \prod_{j \neq i} \nu_j$ with the subscript $-i$ denoting "all components except the ith". Also, let $\Psi_i(\theta; \nu_{-i}) := \mathbb{E}_{-i}[-\log \mathsf{P}(\boldsymbol{x}, \theta)]$. Holding fixed all components ν_{-i} except the ith, the objective function J can be written as a functional of the i-th component as $J_i(\nu_i\,;\,\nu_{-i}) := \int \Psi_i(\theta; \nu_{-i}) \nu_i d\theta_i - H(\nu_i)$. Our formulation and its basic properties motivate the common implementation of MFVI, which takes turns updating each component ν_i^* holding all others fixed. This is called *coordinate ascent variational inference (CAVI)* in the VI literature [2] when the optimization problem is to maximize $-J(\nu)$. Following this approach, we generate a sequence of solutions $\nu_h^k \in \left(\mathcal{P}^2(\mathbb{R}), W_2\right)$

$$\nu_{h,i}^k = \arg\min_{\nu \in \mathcal{P}(\mathbb{R})} \left\{ V_i(\nu\,;\,\nu_h^{k-1}) \;\; := \;\; \frac{1}{2} W_2(\nu_{h,i}^{k-1}, \nu)^2 + h\, J_i(\nu\,;\,\nu_{h,-i}^{k-1}) \right\}. \tag{3}$$

4 MFVI: Gradient Flow Representation

4.1 Gradient Flows on Product Wasserstein Space

Definition 1 (Gradient flow). *A map $\mu(t) \in AC_p(a, b; \prod_{i=1}^d \mathcal{P}_i^2(\mathbb{R}))$ is a solution to the gradient flow equation $j_p(v(t)) \in -\partial\phi(\mu(t))$, if for $v(t) \in T_{\mu(t)}$, its dual vector field $j_p(v(t))$ belongs to the subdifferential of ϕ at μ_t.*

Note that a gradient flow defined here takes the form of $\mu(t) = (\mu_1(t), \mu_2(t), \ldots, \mu_d(t))$, where each $\mu_i(t), i = 1, \ldots, d$ can be viewed as a (marginal) gradient flow in the conventional sense. Meanwhile, the functional $\phi(\mu(t))$ also takes the form $(\,\phi_1(\mu(t)),\, \phi_2(\mu(t)), \ldots, \phi_d(\mu(t))\,)$. While Definition 1 provides a more abstract geometric definition, it can be understood as $\{\mu_i\}_{i=1}^d$ satisfying a system of energy dissipation equations: $\phi_i(\mu(0)) = \phi_i(\mu(t)) + \frac{1}{2}\int_0^t |\dot{\mu}_i((r)|^2 dr + \frac{1}{2}\int_0^t |\nabla\phi_i(\mu(t))|^2 dr$.

Proposition 1. *If the defining functional $\phi(\mu)$ is λ-convex along generalized geodesic for some $\lambda > 0$, then the gradient flow $\mu(t)$ is a contraction, and it is unique given an initial state $\mu(0)$.*

Remark 4.1. *An immediate consequence of Proposition 1 is that the gradient flow $\mu(t)$ of a λ-convex functional geometrically converges to its minimum.*

4.2 Convergence to Gradient Flow on Product Space

Definitions and basic concepts developed in Sec.4.1 enables us to identify the limiting process of the outputs from the discrete algorithm (3) as the step size goes to zero. To be more specific, the limiting gradient flow will take the form of $(\nu_1(t), \nu_2(t), \ldots, \nu_d(t))$ where each $\nu_i(t), i = 1, \ldots, d$ represents the gradient flow in the conventional sense.

Theorem 4.1. *Suppose that the negative log-likelihood function ($-\log \mathsf{P}(x, \theta)$) is λ-convex for some $\lambda > 0$. Define the family of interpolated probability measures $(\nu_{h,1}(t), \nu_{h,2}(t), \ldots, \nu_{h,d}(t))$ for each $k \geq 1$ and $i = 1, \ldots, d$ as $\nu_{h,i}(t) = \nu_{h,i}^k$ for $t \in [kh, (k+1)h)$ where $\{\nu_{h,i}^k\}$ are the updates generated by the discrete algorithm (3). Then, there exists $(\nu_i(t))_{i=1}^d$, a gradient flow on the product space $(\prod_{i=1}^d \mathcal{P}_i^2(\mathbb{R}), W_2)$ defined by a functional $\phi(\nu) = (\phi_1(\nu), \phi_2(\nu), \ldots, \phi_d(\nu))$ with $\phi_i(\nu) = J_i(\nu_i; \nu_{-i})$ for each $i = 1, 2, \ldots, d$, such that, as $h \downarrow 0$, $\nu_{h,i}(t) \rightharpoonup \nu_i(t)$ weakly in W_2 for every $t \in (0, \infty)$.*

Corollary 1. *Suppose that the negative log-likelihood function $-\log \mathsf{P}(\boldsymbol{x}, \theta)$ is λ-convex for some $\lambda > 0$. For $i = 1, \ldots, d$, let $\left\{(\rho_{h,i}^k)_{i=1}^d\right\}_{k \geq 1} \in L^2$ be the densities associated with the measures produced by the iterative scheme (3), and let $\rho_{h,i}(t)$ be their interpolation of $t \in [0, \infty)$ for each h, i. Then, as $h \downarrow 0$,*

$(\rho_{h,1}(t), \rho_{h,2}(t), \ldots, \rho_{h,d}(t)) \rightharpoonup (\rho_1(t), \rho_2(t), \ldots, \rho_d(t))$ *, weakly in* $L^2(\mathbb{R}^d)$ *for a. e.* $t \in (0, \infty)$*, and* $((\rho_1(t), \rho_2(t), \ldots, \rho_d(t)) \in C^\infty((0, \infty) \times \mathbb{R}^d)$ *is the unique solution of the following equation in its coordinate form,*

$$\partial_t \rho_i = \partial_i(\partial_i \Psi_i(x, \rho_{-i})\rho_i) + \partial_i^2 \rho_i, \quad \forall i = 1, \ldots, d \tag{4}$$

with proper initial conditions.

[9] showed first the correspondence between the solution to an evolutional PDE and that of a gradient flow under the condition that the solution is smooth. Recent results under weaker conditions are found in [4] and [5]. Corollary 1 implies that the density of probability measure produced by MFVI algorithm is the solution to the homogeneous version of equation (4), i.e. $\partial_i(\partial_i \Psi_i(x, \rho_{-i})\rho_i) + \partial_i^2 \rho_i = 0, \quad \forall i = 1, \ldots, d$.

4.3 Finite-Time Convergence Rate

Utilize the fact that the gradient flow on the product space approaches the infimum (the target) geometrically, and the error estimation for the discretized gradient flow, i.e. the CAVI algorithm, we can obtain the following convergence rate estimation for MFVI.

Theorem 4.2. *For a* λ*-convex functional* J*, given any* $\epsilon > 0$*, there exists a* $T > 0$ *and* $\eta(T) > 0$*, such that for any* $eta < \eta(T)$*,* $W_2(\nu_\rho^{\lceil T/\eta \rceil}, \nu^*) < \epsilon$*, with* $\lceil T/\eta \rceil$ *denotes the smallest integer that is larger or equal to* T/η*.*

Remark 4.2. *For the problems that the target distribution has* λ *convexity, Theorem 4.2 provides a finite time convergence estimation for MFVI, and this can be viewed as a quantitative extension of Blei & Wang. Furthermore, this can be viewed as a concrete example for our overall pursue of utilizing gradient flows techniques in analyzing general machine learning algorithms.*

5 Uniqueness of quasilinear PDEs

The equation (4) belongs to the following class of quasilinear equations defined on $\mathbb{R}^d \times \mathbb{R}_+$,

$$\partial_t u(x,t) = f(x, u, \nabla_x u) + \triangle_x u(x,t), \qquad u(x,0) = u^0(x) \tag{5}$$

with necessary conditions for $f(x, \xi)$ to be made explicit later.

The existence of weak solutions for quasilinear PDEs has been established in [8]. More specifically, the equation they consider is of the form

$$\begin{cases} \partial_t u - \nabla \cdot (A(t,x)\nabla u) + B(t,x,u,\nabla u) = f, & \text{in } (0,T) \times \Omega, \\ u|_{t=0} = u_0, & \text{in } \Omega, \\ u = 0, & \text{on } (0,T) \times \partial\Omega, \end{cases} \tag{6}$$

where Ω is a regular open bounded set in $\mathbb{R}^d$. Moreover, B has the following form, $B(t,x,u,\nabla u) = b(t,x) \cdot \nabla u + d(t,x)u + g(t,x,u,\nabla u)$. We will utilize the following further assumptions:

Assumption 1. *Boundedness of Coefficients:* $A \in (L^\infty((0,T) \times \Omega))^{d \times d}$*,* $b \in (L^\infty((0,T) \times \Omega))^d$ $d \in L^\infty((0,T) \times \Omega)$*,* $\nabla \cdot b \in L^\infty((0,T) \times \Omega)$*; Uniform Ellipticity: there exists* $a > 0$*, such that,* $A(t,x)\xi \cdot \xi \geq a|\xi|^2$*; The function* $g(t,x,\lambda,\xi) : (0,T) \times \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ *is measurable on* $(0,T) \times \Omega$ *for all* $\lambda \in \mathbb{R}$*,* $\xi \in \mathbb{R}^d$*, and continuous with respect to* λ *and* ξ *almost everywhere in* $(0,T) \times \Omega$*.* $\lambda g(t,x,\lambda,\xi) \geq 0$*, and there exists* $0 \leq \sigma < 2$ *such that* $|g(t,x,\lambda,\xi)| \leq h(|\lambda|)(\gamma(t,x) + |\xi|^\sigma)$ *holds for all* $\lambda \in \mathbb{R}, \xi \in \mathbb{R}^d$*, and almost everywhere in* $(0,T) \times \Omega$*, with* $\gamma \in L^2((0,T) \times \Omega)$*, and* h *a non decreasing function and* ζ*-convex on* $\mathbb{R}^+$ *for some* $\zeta > 0$*;* L^2 *data:* $u_0 \in L^2(\Omega)$*,* $f \in L^2((0,T) \times \Omega)$*.*

Theorem 1 in [8] provides the existence of equation (6), but not the uniqueness when the nonlinear term is not zero, to the best of our knowledge. The following result provides a solid step towards answering the general question.

Theorem 5.1. *Under Assumption 1, the solution to* (6) *is unique.*

6 MFVI: Stochastic Differential Equation Representation

The solution to our quasilinear equation (5) can be viewed as the density function of a weak solution to stochastic differential equation $dX_t = \nabla\nabla\Psi(u(t,(X_t))dt + dw_t$, with $u(t,\cdot)$ denotes the density of $X(t)$ at time t, and w_t a d-dimensional standard Brownian motion, following the arguments in [1]. This equation describes a McKean-Vlasov process. For details on this type of processes, see, e.g. [7]. The output of MFVI corresponds to the stationary distribution of this stochastic process, and thus completes the picture on representations of MFVI. This deep connection opens the doorway to SDE based solutions, a topic of future investigations.

References

[1] V. Barbu and M. Röckner. Probabilistic representation for solutions to nonlinear fokker–planck equations. *SIAM Journal on Mathematical Analysis*, 50(4):4246–4260, 2018.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] V. Bögelein. A variational approach to porous medium type equations. *IMN Internationale Mathematische Nachrichten*, pages 17–32, 2017. 235.

[5] V. Bögelein, F. Duzaar, and P. Marcellini. Existence of evolutionary variational solutions via the calculus of variations. *Journal of Differential Equations*, 256(12):3912–3942, 2014.

[6] R. Carmona, F. Delarue, and A. Lachapelle. Control of mckean–vlasov dynamics versus mean field games. *Mathematics and Financial Economics*, 7(2):131–166, 2013.

[7] T. Funaki. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67(3):331–348, 1984.

[8] T. Goudon and M. Saad. Parabolic equations involving 0th and 1st order terms with l^1 data. *Revista Matemática Iberoamericana*, 17:433–469, 2001.

[9] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[10] M. Lambert, S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet. Variational inference via wasserstein gradient flows. *arXiv:2205.15902*, 2022.

[11] N. G. Trillos and D. Sanz-Alonso. The Bayesian Update: Variational Formulations and Gradient Flows. *Bayesian Analysis*, 15(1):29 – 56, 2020.

[12] Y. Wang and D. M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

[13] R. Yao and Y. Yang. Mean field variational inference via wasserstein gradient flow, 2022.