THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A
Volume 23, Number 2/2022, pp. 107–111

VARIATIONAL BAYESIAN APPROXIMATION. A RIGOROUS APPROACH

Alireza BAHRAINI

Sharif University of Technology, Department of Mathematical Sciences
P.O.Box 11155-9415, Tehran, Iran
Corresponding author: Alireza BAHRAINI, E-mail: bahraini@sharif.edu

Abstract. We apply the theory of optimal transport to study mathematical properties of mean field variational Bayesian approximation. It turns out that if $K+C>0$ where C is the convexity coefficient of $-\log p$ and K is a lower bound for the Ricci curvature of the underlying parameter space, then the corresponding system of equations of variational Bayesian approximation admits a unique solution. The uniqueness property in presence of symmetry leads to preservation of mode. As an explicit application we correct Bayesian Gaussian Mixture model in such a way that it turns into a convex model while its (unique) maximum likelihood solution coincides asymptotically with the true solution. Using convexity it is possible to prove asymptotic accuracy of the mode obtained by mean field variational Bayesian approximation. This seems to be the first rigorous proof for this fundamental fact which was expected based on several experimental computations.

Key words: MFVBA, pptimal transport, Ricci curvature, Gaussian mixture model, geodesic convexity

1. INTRODUCTION

Let $\mathscr{M}(\mathscr{H})$ denote the space of Borel probability measures associated to a given Polish space $\mathscr{H}$. If $\mathscr{H}=\Pi_{i=1}^{K}\mathscr{H}_i$ where $\mathscr{H}_i$ for $i=1,...,K$ are Polish spaces the minimization problem

$$\arg\min_{\nu\in\mathscr{A}} D_{KL}(\nu||\mu) \tag{1}$$

where

$$\mathscr{A}:=\Pi_{i=1}^{K}\mathscr{M}(\mathscr{H}_i)$$

is called Mean Field Variational Bayesian Approximation (MFVBA) of μ by a factorized probability measure ν. Here D_{KL} is the Kullback-Leibler Divergence (KLD) defined on $\mathscr{M}(\mathscr{H})$ as follows:

$$D_{KL}(\nu||\mu)=\int_{\mathscr{H}}\log(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x))\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\mu(\mathrm{d}x)=\mathbb{E}^{\mu}[\log(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x))\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)].$$

The KLD for not absolutely continuous measures is set to be equal to ∞. The existence of solution to this variational problem can simply be established by weak convergence approach [3] but the solution in general is not unique.

MFVBA has been applied as a powerful tool for parameter estimation in graphical models in analyzing large scale data such as Dimensionality Reduction (DR), Principle Component Analysis (PCA), Independent Component Analysis (ICA), Clustering and Classification using mixture models [6], meanwhile we do not yet know any mathematically rigorous framework justifying, measuring and estimating the range of applicability and error bounds of this Approximate Bayesian Method (see e.g. [2] and [1]). In this note we apply the connection established by Lott-Villani-Sturm [4] and [5] between the theory of optimal transport and Ricci curvature to study convexity properties of the relative entropy in order to find conditions under which MFVBA admits a unique solution. We then apply this observation to correct Gaussian mixture model in such a way that the convergence of the corresponding maximum likelihood solution towards the right solution (with growth of data) is preserved and moreover the new model turns into a convex one. This will guarantee the convergence of MFVBA towards the expected solution in appropriate asymptotic parameter regimes.

We recall that, using variational method, the solution of the optimization problem (1) can be described by the following system of integral equations:

$$\log(\frac{\mathrm{d}\nu_i}{\mathrm{d}\omega_i}(x_i)) = \mathbb{E}^{\nu_{\setminus i}}(\log\frac{\mathrm{d}\mu}{\mathrm{d}\omega}) \tag{2}$$

where $\omega = \Pi_i \omega_i \in \mathscr{M}(\mathscr{H})$ is an appropriate reference measure for some $\omega_i \in \mathscr{M}(\mathscr{H}_i)$ and $\nu_{\setminus i} = \Pi_{j\neq i}\nu_j$. The existence of solutions for this system of mean field type equations can be established under weak constrains on $\mathscr{A}$ although the uniqueness does not hold in general at all.

2. CONVEXITY OF KULLBACK-LEIBLER FUNCTIONAL: CONTINUOUS VERSUS DISCRETE CASE

2.1. Continuous parameter space

Assume that the Polish space $\mathscr{H}$ is a complete Riemannian manifold which admits a splitting as $(\mathscr{H} = \Pi_{i=1}^N \mathscr{H}_i, g = \Pi_{i=1}^N g_i)$ into a finite number of complete Riemannian sub-manifolds $(\mathscr{H}_i, g_i)$ for $i = 1,..,N$ where g_i, $i = 1,...,N$ is a Riemannian metric on $\mathscr{H}_i$. We also assume that there exists a lower bound K for the Ricci curvature of all the spaces $(\mathscr{H}_i, g_i)$.

Let $\mathscr{P}_2(\mathscr{H})$ be the space of probability measures on $\mathscr{H}$ which are absolutely continuous with respect to ω_g and such that $\frac{d\mu}{d\omega_g} \in L^2(\mathscr{H}, \omega_g)$. Here ω_g denotes the volume form associated to g. Consider a probability measure $\mu \in \mathscr{P}_2(\mathscr{H})$ defined as

$$\frac{\mathrm{d}\mu}{\mathrm{d}\omega}(x) = \frac{1}{Z_\mu}\exp(-\Phi(x)), \tag{3}$$

where $\Phi : \mathscr{H} \to \mathbb{R}$ denotes a map on $\mathscr{H}$ and where and Z_μ is the normalization constant.

The application $\Phi : \mathscr{H} \to \mathbb{R}$ is assumed to be a C-convex function with respect to the geometry of $\mathscr{H}$ induced by g, where C is a real constant.

The proof of the following proposition is straightforward:

PROPOSITION 1. *Let $\mu \in \mathscr{P}_2(\mathscr{H})$ be a measure with all the above mentioned properties and let $\mathscr{A} = \Pi_i \mathscr{H}_i$. Let ν_1 and ν_2 be two separable measures with compact support belonging to $\mathscr{A}$, such that $D_{KL}(\nu_1||\mu) < \infty$ and $D_{KL}(\nu_2||\mu) < \infty$. Consider an interpolating measure $\nu_t^{1\to 2} = \Pi_i \nu_i^t \in \mathscr{A}$, for $0 \le t \le 1$ which is a Wasserstein geodesic on each component $\nu_i^t \in \mathscr{M}(\mathscr{H}_i)$. Then we have*

$$D_{KL}(\nu_t^{1\to 2}||\mu) \le (1-t)D_{KL}(\nu_1||\mu) + tD_{KL}(\nu_1||\mu) - \frac{t(1-t)}{2}(K+C)W_2(\nu_0,\nu_1)^2, \tag{4}$$

where K is the common lower bound for Ricci curvature of all the spaces $\mathscr{H}_i$ and C is the coefficient of convexity of Φ. In particular if $C+K > 0$ then D_{KL} is convex on $\mathscr{A}$ and thus the solution of the optimization problem $\operatorname{argmin}_{\nu\in\mathscr{A}} D_{KL}(\nu||\mu)$ *is unique.*

COROLLARY 1. *In the case where a group G acts on each of the factors $\mathscr{H}_i$'s with a single fixed point occurring at the unique maximum of $\frac{\mathrm{d}\mu}{\mathrm{d}\omega}$ then the mode of MFVBA of μ coincides with the mode of μ.*

Example. In the case where μ is a Gaussian measure on Euclidean space $\mathbb{R}^n$ the maximum of its MFVBA is the same of the maximum of μ.

In many applications of the MFVBA, like its application for the GMM, the splitting of the space of random variables $\mathscr{H}$ is of the form $\mathscr{H}_1 \times \mathscr{H}_2$, where one of the factors for instance $\mathscr{H}_1$ is a Riemannian manifold, but the other factor $\mathscr{H}_2$ is a discrete space. We study the convexity of this case in the following section.

2.2. A mixed discrete-continuous version

Let the Polish space $\mathscr{H}$ be written as a product $\mathscr{H} = \mathscr{H}_1 \times \mathscr{H}_2$ in which $\mathscr{H}_1$ is a complete Riemannian manifold equipped with a Riemannian metric g while $\mathscr{H}_2$ is a possibly finite set. This case occurs in the MFVBA of Gaussian Mixture Model which is the model we want to study using theory of optimal transport.

Let $(\mathscr{H}_{1,\lambda}, g_\lambda)$, for $\lambda > 0$, denote the Riemannian manifold obtained from $\mathscr{H}_1$ through a rescaling of the metric g as $g_\lambda := \frac{1}{\lambda} g$.

Then we can prove the following theorem:

THEOREM 1. *Let K denote a lower bound for the Ricci curvature of $(\mathscr{H}_1, g)$ and let $\Phi = -\log(\frac{d\mu}{d\omega_g})$ be C-convex when restricted to each connectivity component $\mathscr{H}^i := \mathscr{H}_1 \times \{i\}$ for $i \in \mathscr{H}_2$ where C is independent of i. Also assume that $K + C > 0$. Then for large values of λ the Kullback-Leibler functional $\nu \to D_{KL}(\nu||\mu)$ will be a convex functional over $\mathscr{A}$ and thus the solution to the optimization problem $argmin_{\nu \in \mathscr{A}} D_{KL}(\nu||\mu)$ is unique. Here $\mathscr{A}$ is the subspace of factorized probability measures on $\mathscr{H}$ as in VBA.*

Proof. First we have to explain in what sense we are talking about the convexity here. More precisely what are the class of geodesics we are considering on the space of probability measures on $\mathscr{P}_2(\mathscr{H}_1) \times \mathscr{M}(\mathscr{H}_2)$.

We consider a separable geodesic path of probability measures $\nu^{1\to 2}_{t,\lambda g} := \nu^1_{t,\lambda g} \times \nu^2_t$ on $\mathscr{P}_2(\mathscr{H}_1) \times \mathscr{M}(\mathscr{H}_2)$ as being the product of a geodesic ν^1_t over $\mathscr{P}_2(\mathscr{H}_1)$ in the sense of its natural Wasserstein structure as discussed before, and a path $\nu^2_t(i_1, ..., i_K) := \Pi^K_{j=1}((1-t)a_{ij} + tb_{ij})$ joining two probability measures ν^2_0 and ν^2_1 on $\mathscr{M}(\mathscr{H}_2)$. In fact we are assuming that $\mathscr{H}_2 = \{i_1, ..., i_K\}$ and

$$(\nu^2_0(i_1), ..., \nu^2_0(i_K)) = (a_{i_1}, ..., a_{i_K}) \qquad (\nu^2_1(i_1), ..., \nu^2_1(i_K)) = (b_{i_1}, ..., b_{i_K}).$$

Now if we rescale the metric g on the underlying space $\mathscr{H}_1$ as $g_\lambda := \frac{1}{\lambda} g$ the impact will appear on the geodesic $\nu^{1\to 2}_t$ like $\nu^{1\to 2}_{t,\frac{1}{\lambda}g} = \nu^{1\to 2}_{\lambda t, g}$. Also the curvature tensor and in the case of metric measure spaces the lower Ricci bound will scale like $\mathrm{Ricci}(\frac{1}{\lambda} g) = \lambda^2 \mathrm{Ricci}(g)$.

It is not difficult to see that we can restrict the above mentioned geodesic paths to those with compact support on $\mathscr{H}_1$. Now we can write:

$$\begin{aligned}
\frac{d^2}{dt^2}\mathbb{E}^{\nu^{1\to 2}_{t,\lambda g}}(\Phi(x)) &= \frac{d^2}{dt^2}\int_y (\sum_z \nu^2_t(z)\Phi(z,y))\nu^1_{t,\lambda g}(y) \\
&= \int_y (\sum_z \frac{d^2\nu^2_t(z)}{dt^2}\Phi(z,y))\nu^1_{t,\lambda g}(y) + \int_y (\sum_z \nu^2_t(z)\Phi(z,y))\frac{d^2\nu^1_{t,\lambda g}(y)}{dt^2} \\
&\quad + \int_y (\sum_z \frac{d\nu^2_t(z)}{dt}\Phi(z,y))\frac{d\nu^1_{t,\lambda g}(y)}{dt} \\
&= \int_y (\sum_z \frac{d^2\nu^2_t(z)}{dt^2}\Phi(z,y))\nu^1_{t,\lambda g}(y) + \int_y (\sum_z \nu^2_t(z)\frac{d^2}{dt^2}\Phi(z, G_{t,\lambda g}(y)))\nu^1_{0,\lambda g}(y) \\
&\quad + \int_y (\sum_z \frac{d\nu^2_t(z)}{dt}\frac{d}{dt}\Phi(z, G_{t,\lambda g}(y))))\nu^1_{0,\lambda g}(y) \\
&= \int_y (\sum_z \frac{d^2\nu^2_t(z)}{dt^2}\Phi(z,y))\nu^1_{t,\lambda g}(y) + \int_y (\sum_z \nu^2_t(z)\frac{d^2}{dt^2}\Phi(z, G_{\frac{t}{\lambda},g}(y)))\nu^1_{0,\lambda g}(y) \\
&\quad + \int_y (\sum_z \frac{d\nu^2_t(z)}{dt}\frac{d}{dt}\Phi(z, G_{\frac{t}{\lambda},g}(y))))\nu^1_{0,\lambda g}(y)
\end{aligned}$$

$$= \int_y (\sum_z \frac{\mathrm{d}^2 \nu_t^2(z)}{\mathrm{d}t^2} \Phi(z,y)) \nu^1_{t,\lambda g}(y) + \frac{1}{\lambda^2} \int_y (\sum_z \nu_t^2(z) \frac{\mathrm{d}^2 \Phi(z, G_{t,g}(y))}{\mathrm{d}t^2})|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$$
$$+ \frac{1}{\lambda} \int_y (\sum_z \frac{\mathrm{d}\nu_t^2(z)}{\mathrm{d}t} \frac{\mathrm{d}\Phi(z, G_{t,g}(y)))}{\mathrm{d}t})|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$$

$$G_{t,\lambda g}(y) = \exp^{\lambda g}\{-t \nabla^{\lambda g} F\}.$$

Then since $\nabla^{\lambda g} F = \frac{1}{\lambda} \nabla^g F$ and $\exp^{\lambda g} = \exp^g$ we get

$$G_{t,\lambda g}(y) = G_{\frac{t}{\lambda},g}(y).$$

According to the hypothesis on C convexity of Φ we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathbb{E}^{\nu^{1\to 2}_{t,\lambda g}}(\Phi(x)) = \frac{A}{\lambda^2} + \frac{B}{\lambda},$$

where $A = \int_y (\sum_z \nu_t^2(z) \frac{\mathrm{d}^2 \Phi(z,G_{t,g}(y))}{\mathrm{d}t^2})|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$ and $B := \int_y (\sum_z \frac{\mathrm{d}\nu_t^2(z)}{\mathrm{d}t} \frac{\mathrm{d}\Phi(z,G_{t,g}(y)))}{\mathrm{d}t})|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$.

$$\frac{\mathrm{d}^2 \Phi(z, G_{t,g}(y))}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t} < \nabla^g_y \Phi(z, G_{t,g}(y)), \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t} > =$$
$$= \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}^T \mathrm{Hess}_{g,y}(\Phi) \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}$$

It follows that

$$A = \int_y \sum_z \nu_t^2(z) (\frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}^T \mathrm{Hess}_{g,y}(\Phi(z, G_{t,g}(y))) \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t})|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$$

$$B = \int_y (\sum_z \frac{\mathrm{d}\nu_t^2(z)}{\mathrm{d}t} < \nabla^g_y \Phi(z, G_{t,g}(y)), \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t} >)|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$$

$$A \geq C \int_y (|\frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}|^2)|_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y) = C \int_y (|\frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}|^2)|_{t=0} \nu^1_{0,\lambda g}(y).$$

The last equality holds because G_t is a geodesic and so its velocity has a constant norm.

$$|B| \geq - \int_y (\sum_z \frac{\mathrm{d}\nu_t^2(z)}{\mathrm{d}t} |\nabla^g_y \Phi(z, G_{t,g}(y))|^2 | \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t})|^2 |_{\frac{t}{\lambda}} \nu^1_{0,\lambda g}(y)$$
$$= - \int_y (\sum_z \frac{\mathrm{d}\nu_t^2(z)}{\mathrm{d}t} |\nabla^g_y \Phi(z, G_{t,g}(y))|^2 | \frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t})|^2 |_{t=0} \nu^1_{0,\lambda g}(y)$$
$$\geq C' \int_y (|\frac{\mathrm{d}G_{t,g}(y)}{\mathrm{d}t}|^2)|_{t=0} \nu^1_{0,\lambda g}(y)$$

where C' depends on Φ and $\mathscr{H}$.

For the entropy functional $H(\nu^{1\to 2}_{t,\lambda g})$ of the path of separable probability measure $\nu^{1\to 2}_{t,\lambda g}$ we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} H(\nu^{1\to 2}_{t,\lambda g}) \geq \frac{1}{\lambda^2} K + C''$$

which is due to the fact that after change of scale in metric as $g \to \lambda g$ the Ricci curvature scales as

$\mathrm{Ricci}(\lambda g) = \frac{1}{\lambda^2}\mathrm{Ricci}(g)$, thus we get

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} D_{KL}(\nu^{1\to 2}_{t,\lambda g}||\mu) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{E}^{\nu^{1\to 2}_{t,\lambda g}}(\Phi(x)) + \frac{\mathrm{d}^2}{\mathrm{d}t^2} H(\nu^{1\to 2}_{t,\lambda g}) \geq (C+K)\frac{1}{\lambda^2} + C'\frac{1}{\lambda} + C''$$

since we know that $K+C>0$ then for small enough λ the right hand side of the above inequality will become positive hence convexity of D_{KL} along these geodesics follows.

Example (Correction to GMM). We introduce the probability density $P_{N,new,\lambda}$ whose logarithm is defined by

$$-\log P_{N,new,\lambda} = \lambda(-\log P_N + \sum_k \theta_k \|\mu_k - \bar{x}_k\|^4 + \frac{1}{2}\sum_k\sum_i \eta_k z_{ik}(x_i - \bar{x}_k)^T \Lambda_k (x_i - \bar{x}_k) - \frac{1}{2}\eta_k \log|\Lambda_k|) + \mathrm{const}.$$

in which $\bar{x}_k := \frac{\sum_i z_{ik} x_i}{N_k}$, and the log posterior $\log P_N$ is given by

$$\begin{aligned} \log P_N(z,\mu,\pi,\Lambda|x) &= \textstyle\sum_{n=1}^N \sum_{k=1}^K z_{nk}(\log \pi_k) - \frac{1}{2}(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2}\log|\Lambda_k|) + \\ &\quad + \textstyle\sum_{k=1}^K \log p(\mu_k) + \sum_{k=1}^K \log p(\Lambda_k) + \log p(\pi) + C \end{aligned} \tag{5}$$

(see page 6 of [1] for details of notations).

Here the space $\mathscr{H}_1$ contains the continuous parameters μ_k, Λ_k, π_k and $\mathscr{H}_2 = \{1,...,K\}^N$ where N is the number of data points and K the number of classes of the data.

There exists a natural Riemannian metric on $\mathscr{H}_1$ such that this metric and the probability $P_{N,new,\lambda}$ for large values of λ satisfy the hypothesis of theorem (1). Thus for large values of λ, $-\log P_{N,new,\lambda}$ admits a unique minimum. Also the minimum of $-\log P_{N,new,\lambda}$ tends to the absolute minimum of $-\log P_N$ as $N \to +\infty$. The convergence of the unique mode of the MFVBA of $P_{N,new,\lambda}$ to the mode of $P_{N,ew,\lambda}$ results from these two observations.

ACKNOWLEDGEMENTS

The author is grateful to the research council of Sharif University of Technology for support.

REFERENCES

1. J.R. GIORDANO., T. BRODERICK, M.I. JORDAN, *Linear response methods for accurate covariance estimates from mean field variational Bayes*, Neural Information Processing Systems Conference (NIPS), 2015.
2. A.Y. ZHANG, H.H. ZHOU, *Theoretical and computational guarantees of mean field variational inference for community detection*, The Annals of Statistics, **48**, *5*, pp. 2575–2598, 2020.
3. P. DUPUIS, R.S. ELLIS, *A weak convergence approach to the theory of large deviations*, John Wiley & Sons, Inc., New York, 1997.
4. J. LOTT, C. VILLANI, *Ricci curvature for metric-measure spaces via optimal transport*, Annals of Mathematics, **169**, *3*, pp. 903–991, 2009.
5. K.-T. STURM, *On the geometry of metric measure spaces*, Acta Mathematica, **196**, *1*, pp. 65–131, 2006.
6. D.M. BLEI, A. KURCUKELBIR, J.D. McAULIFFE, *Variational inference: A review for statisticians*, Journal of the American Statistical Association, **112**, *518*, pp. 859–877, 2017.

Received November 6, 2021