Black Box Variational Bayesian Model Averaging

Shrijita Bhattacharya

Michigan State University

December 14, 2021

Acknowledgments and references

Parts of the talk are based on the joint work with

- Dr. Taps Maiti (MSU)
- Dr. Vojtech Kejzlar (Skidmore College)
- Mookyong Son (MSU)

Main references:

- Vojtech Kejzlar, Shrijita Bhattacharya, Mookyong Son, Tapabrata Maiti. *Black Box Variational Bayes Model Averaging*. Preprint: arXiv:2106.12652

Table of Contents

Bayesian Inference

- Quantity of interest: Δ - unknown parameters $\boldsymbol{\theta}$ or new observation y_{new}

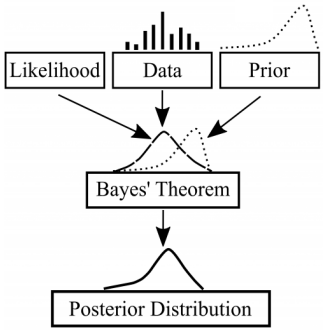
Posterior distribution

Data likelihood

Prior distribution

$$P(\Delta|y) = \frac{P(y|\Delta)P(\Delta)}{P(y)}$$

Marginal likelihood



courtesy:jason-doll.com/wordpress/?page_id=127

Our main interest here is the posterior $p(\boldsymbol{\theta}|y)$

Bayesian Inference

MCMC methods are popular to approximate $p(\boldsymbol{\theta}|\boldsymbol{y})$:

- Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995)
- No-U-Turn sampler (NUTS) (Homan and Gelman, 2014)

- **Become quickly impractical with the increasing size of datasets, number of parameters, and model complexity**

Solution: Variational (Bayesian) Inference

- Approximation of target density through optimization

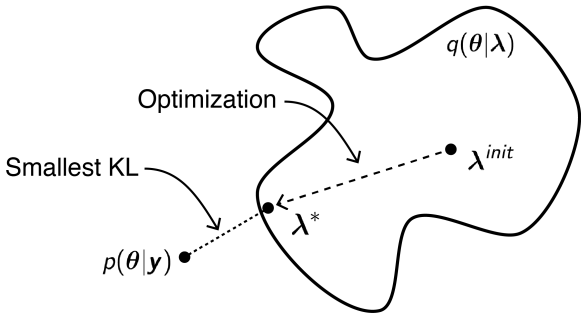
Table of Contents

Variational Bayes Inference - Overview

- VBI approximates $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a family of distributions $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ indexed by parameter $\boldsymbol{\lambda}$

$$q^* = \underset{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\arg\min}\, KL(q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\theta|\boldsymbol{y})).$$

- $KL(q||p) = \mathbb{E}_q\left[\log \frac{q}{p}\right]$ "measures" the loss of information when p is approximated by q



Variational Bayes Inference - Overview

KL is intractable, in practice we maximize the *evidence lower bound (ELBO):*

$$\mathcal{L}(\boldsymbol{\lambda}) = \underbrace{\mathbb{E}_q\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right]}_{\text{Expected log-likelihood of data}} - \underbrace{KL(q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}))}_{\text{KL between variational and prior}} .$$

- First term prefers q to place mass on the MLE
- Second term encourages q to be close to the prior
- **The ELBO is a lower bound on log evidence:** $\mathcal{L}(\boldsymbol{\lambda}) \leq \log p(\boldsymbol{y})$

Variational Bayes Inference - Overview

- ELBO can be maximized by any optimization technique
- The scalability of VBI is achieved through stochastic gradient ascent (SGA)

SGA

Let $\tilde{l}(\boldsymbol{\lambda})$ be a realization of the random variable $\tilde{\mathcal{L}}(\boldsymbol{\lambda})$, so that $\mathbb{E}(\tilde{\mathcal{L}}(\boldsymbol{\lambda})) = \nabla_{\lambda}\mathcal{L}(\boldsymbol{\lambda})$. SGA updates $\boldsymbol{\lambda}$ at the t^{th} iteration with

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \rho_t \tilde{l}(\boldsymbol{\lambda}_t).$$

SGA converges to a local maximum of $\mathcal{L}(\boldsymbol{\lambda})$ (global for $\mathcal{L}(\boldsymbol{\lambda})$ concave (Bottou et al., 1997)) when the learning rate ρ_t follows the Robbins-Monro conditions (Robbins and Monro, 1951)

$$\sum_{t=1}^{\infty} \rho_t = \infty, \qquad \sum_{t=1}^{\infty} \rho_t^2 < \infty.$$

Variational Bayes Inference - Implementation Details

Variational Families:

- Mean field family - components of $\boldsymbol{\theta}$ are assumed to be independent

$$q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = q(\theta_1|\boldsymbol{\lambda}_1) \times \cdots \times q(\theta_k|\boldsymbol{\lambda}_k)$$

- Hierarchical families (Ranganath et al., 2016)
- Normalizing flows (Papamakarios et al., 2021)

Learning Rate:

- Adaptive learning rates: AdaGrad (Duchi et al., 2011), Adam (Kingma and Ba, 2014) , RMSProp (Tieleman and Hinton, 2012)

Variational Bayes Inference - Examples

What is VBI good for:

- Stochastic optimization scales up VBI to massive dataset
- The Black-box gradients generalize VBI to a wide class of models

Generalized linear models Bayesian neural networks

Deep exponential families Gaussian processes

- Bayesian Model Averaging / Model Mixing

  EX: We have more than one model of binding energies

Table of Contents

Bayesian Model Mixing - Overview

- **Multiple models competing models to solve the same or similar problems**
- For any quantity of interest $\boldsymbol{\Delta}$, the BMM posterior density $p(\boldsymbol{\Delta}|\boldsymbol{y})$ corresponds to

$$p(\boldsymbol{\Delta}|\boldsymbol{y}) = \sum_{M \in \mathcal{M}} p(\boldsymbol{\Delta}|\boldsymbol{y}, M) p(M|\boldsymbol{y}),$$

where $\mathcal{M}$ denotes the space of all models and

$$p(M|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M)p(M)}{\sum_{M' \in \mathcal{M}} p(\boldsymbol{y}|M')p(M')},$$

and $p(\boldsymbol{y}|M)$ is the model evidence

$$p(\boldsymbol{y}|M) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_M, M) p(\boldsymbol{\theta}_M|M) \, \mathrm{d}\boldsymbol{\theta}_M,$$

Bayesian Model Mixing - VBI approach

The VBI solution to BMM is formulated as

$$q^* = \arg\min_q KL(q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M)||p(M, \boldsymbol{\theta}_M|\boldsymbol{y})),$$

where the variational approximation of the joint posterior $p(M, \boldsymbol{\theta}_M|\boldsymbol{y})$ is of the form

$$q(M, \boldsymbol{\theta}_M|\boldsymbol{\lambda}_M) = q(M)q(\boldsymbol{\theta}_M|M, \boldsymbol{\lambda}_M).$$

- M is assumed to be a random variable whose values are the individual models in the model space $\mathcal{M}$
- $\boldsymbol{\lambda}_M$ is the variational parameter of the family for $\boldsymbol{\theta}_M$ under the model M

Bayesian Model Mixing - VBI approach

The VBI algorithm for BMM:

1. Update $\boldsymbol{\lambda}_M$ as

$$\boldsymbol{\lambda}_M^{t+1} = \boldsymbol{\lambda}_M^t + \rho q(M) \tilde{l}(\boldsymbol{\lambda}_M^t),$$

where $\tilde{l}(\boldsymbol{\lambda}_M)$ is a MC estimate of ELBO gradient of model M.

2. Update $q(M)$ as

$$q(M) \propto \exp(\widehat{\mathcal{L}}_M + \log p(M)),$$

where $\widehat{\mathcal{L}}_M$ is a MC estimate of ELBO of model M.

- Enjoys all the advantages of VBI (scalability, generalizibility)
- Model weights estimates $q(M)$ are fairly stable
- See (Kejzlar et al., 2021) for details

Bayesian Model Mixing - VBI approach - Example

Aggregated crime data on 47 U.S. states of Vandaele (1978).

Given the crime rate y, we consider models of the form

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

- $x_1, \ldots, x_p$ is a subset of a set of candidate predictors $x_1, \ldots, x_k$
- For simplicity, we only consider 2 candidate predictors

x_1	percentage of males age 14-24
x_2	probability of imprisonment

Bayesian Model Mixing - Example

	Inclusion			$p(M|\boldsymbol{y})$	
Model	**Intercept**	x_1	x_2	**MC**	**VBI**
0	*			0.03	0.06
1	*		*	0.78	0.78
2	*	*		< 0.01	< 0.01
3	*	*	*	0.19	0.15

- Zellner's g-prior (Zellner, 1986) was used for model parameters
- Mean-field approximation was used for model parameters
- For more examples, see (Kejzlar et al., 2021)

Bayesian Model Mixing - VBI approach - Example

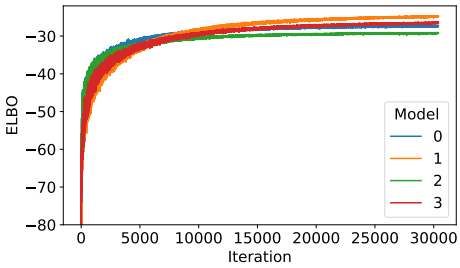


Figure: Converging ELBOs for all the 4 models on the set of two predictors.

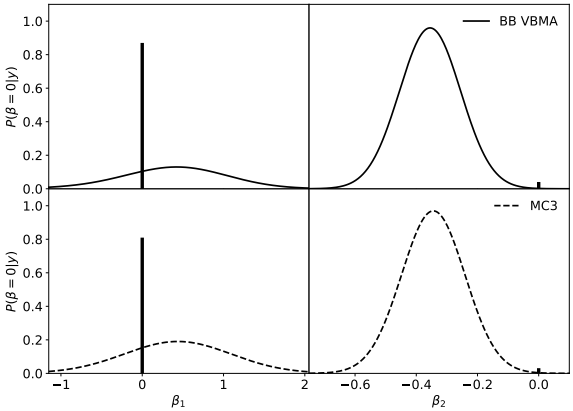


Figure: Posterior distributions for β_1 and β_2 based on MC3 and BBVBMA.

Bayesian Model Mixing - VBI approach - Example

Table: Predictive coverage of the test set based on the 90% predictive credible intervals. Model averaging is compared with the performance of the models chosen by two model selection method (Adjusted R^2 and Mallows' C_p).

		Predictive coverage (%)	
Method	**Model**	**MC**	**BVI**
Model averaging		0.88	0.88
Adjusted R^2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	0.84	0.84
Mallows' C_p	$\beta_0 + \beta_2 x_2$	0.84	0.84

Table of Contents

References

Bottou, L., Le Cun, Y., and Bengio, Y. (1997). Global training of document processing systems using graph transformer networks. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 489–493. IEEE.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Homan, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381.

Kejzlar, V., Bhattacharya, S., Son, M., and Maiti, T. (2021). Black box variational bayes model averaging. *ArXiv:2106.12652*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Ranganath, R., Tran, D., and Blei, D. M. (2016). Hierarchical variational models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48*, ICML'16, pages 2568–2577. JMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Vandaele, W. (1978). Participation in illegitimate activities-ehrlich revisited (from deterrence and incapacitation-estimating the effects of criminal sanctions on crime rates, p 270-335, 1978, alfred blumstein et al, ed.-see ncj-44669).

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Brune de Finetti*, pages 233–243.