Extending Variational Boosting to Multiple Competing Variational Families

Naitong Chen

April 24, 2021

1 Introduction

In recent years, Bayesian statistical models have received a lot of attention thanks to their ability to model data of complex structures, incorporate domain expertise, and coherently quantify uncertainties around the estimates of the latent parameters. One of the key challenges of applying Bayesian statistical models in practice is to obtain samples from the often intractable posterior distribution. Markov chain Monte Carlo (MCMC) and Variational Inference (VI) are the two main approaches that address this challenge. MCMC generates posterior samples by simulating a Markov chain whose invariant distribution is the target posterior distribution. One of the main advantage of MCMC is that there is a tradeoff between computation time and the quality of the posterior samples: as more samples are generated by simulating the Markov chain for longer time, one can expect the samples to be better approximate the posterior distribution. However, MCMC does not scale well to problems with large datasets. Specifically, due to the iterative nature of MCMC, the log likelihood of each observation in the entire dataset needs to be evaluated before a single posterior sample can be generated. VI handles this scalability issue by approximating the target using some tractable distribution. More precisely, VI finds some distribution from a tractable distribution family that minimizes some divergence to the target distribution. The tractablity of the posterior approximation allows us to easily sample from the posterior distribution. However, the quality of the posterior approximation under the VI approach is fundamentally limited by the choice of the variational family. In other words, the approximation does not get closer to the true target distribution with more computation if the variational family is poorly chosen.

Inspired by the idea of boosting, the Variational Boosting (VB) algorithm proposed in [MFA17] addressed this shortcoming of the VI methods. Specifically, VB builds a mixture approximation of the target by interatively adding components from some variational family. This approach not only increases the expressibility of a fixed variational family by extending the variational family to its convex hull, but also forms the tradeoff between computation time and the quality of approximation, just like MCMC. A later paper then

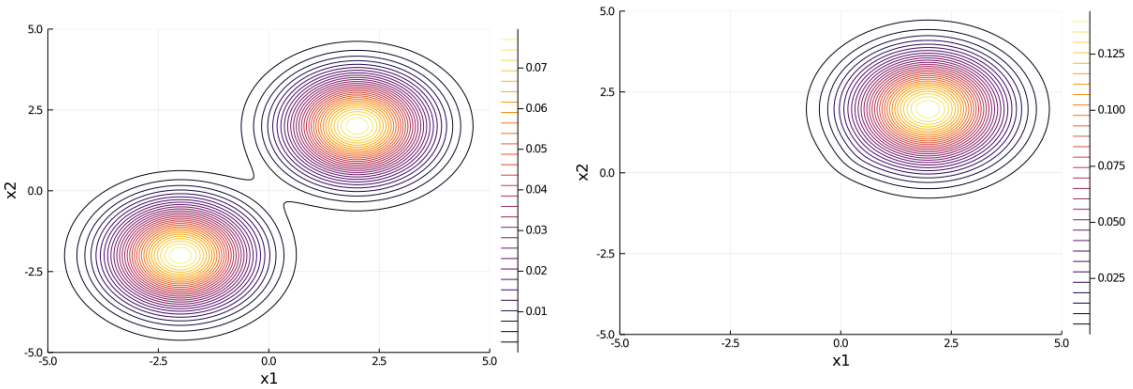
confirmed that under certain conditions, as the number of components increased, the mixture approximation from VB would indeed converge to the true target distribution [Loc+18].

We note that the work in [MFA17] focused on a single variational family. Although it has been shown that the mixture approximation converges to the target distribution as the number of mixture components increase, the choice of variational family may still result in varying efficiencies of the algorithm. An example is to approximate a heavy-tailed target distribution using mixtures of Gaussian components. The light-tailedness of the Gaussian distribution may hinder the rate of convergence in terms of the number of components, requiring many mixture components to obtain reasonable approximations of the target distribution's tail behaviour. One possible way to address this limitation is to further diversify the variational family by considering multiple candidate components from different distribution families at each iteration. More concretely, at each iteration, instead of optimizing the next component to be added to the mixture approximation from a single distribution family, multiple candidate components from different distribution families can be optimized at the same time. Then the component that yields the mixture approximation with the smallest divergence to the target distribution is added.

In this report, we begin in Section 2 by reviewing the VB algorithm and two practical considerations that improve the efficiency of the method introduced in [MFA17]. Experiments conducted on synthetic data are shown in Section 3 to verify and critique the effectiveness of the VB algorithm. We then describe our proposed extension and analyze some experiment results on some synthetic data in Sections 4 and 5.

2 Variational Boosting

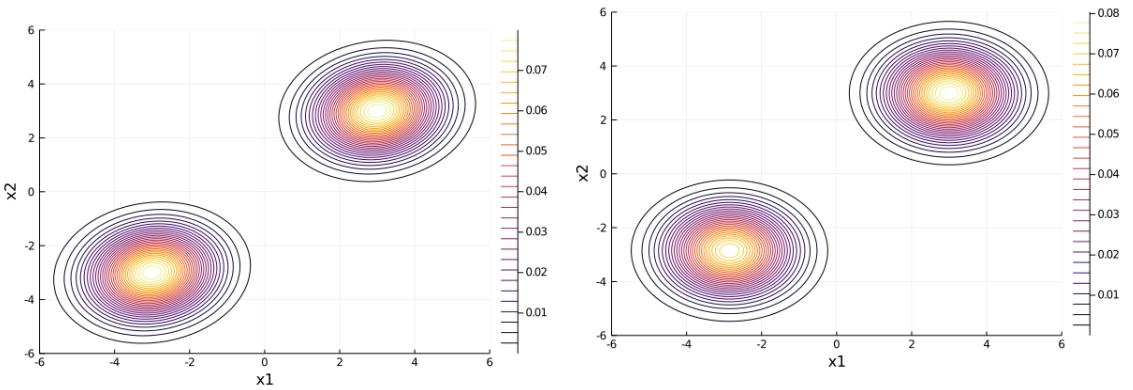
3 Verifying Variational Boosting's Effectiveness



(a) True posterior distritbution

(b) Variational boosting approximating with two components

Figure 1: Comparing the target distribution and the variational boosting approximation using the weighed EM component initialization.



(a) True posterior distritbution

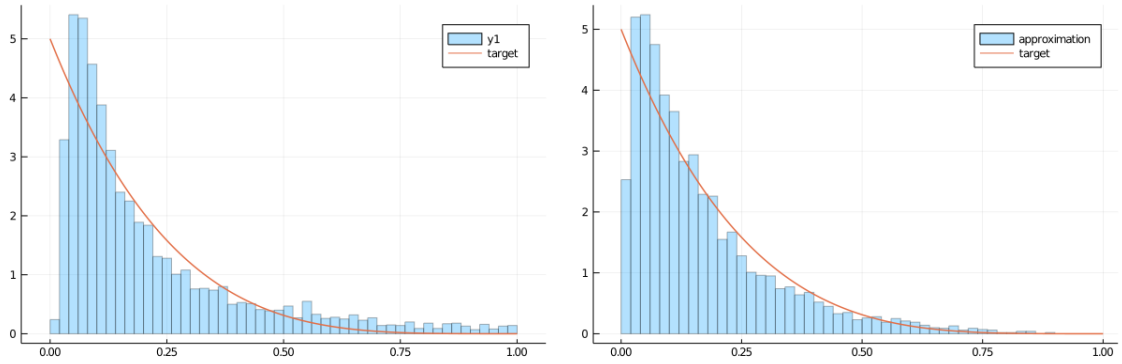
(b) Variational boosting approximating with two components

Figure 2: Comparing the target distribution and the variational boosting approximation without the weighed EM component initialization.

4 Extending to Multiple Variational Families

C [KW13]

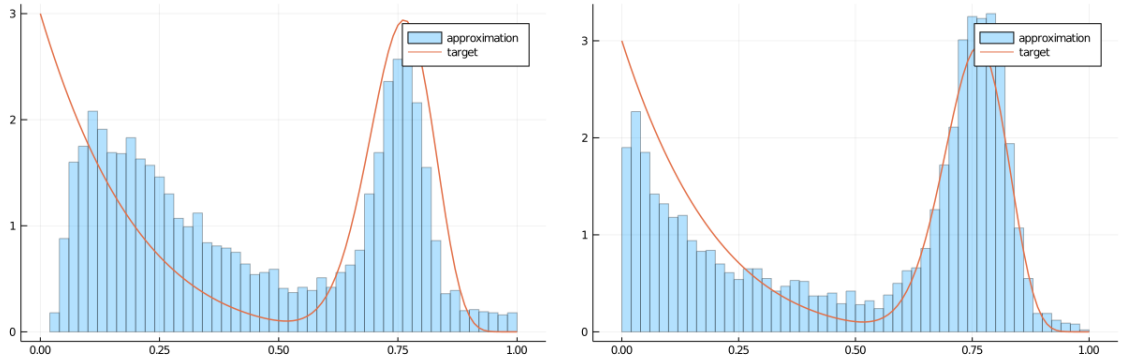
5 Experiments



(a) Approximating Beta$(1, 5)$ using a single exponential component

(b) Approximating Beta$(1, 5)$ using a single log-normal component

Figure 3: Comparing the single component approximations on a Beta distribution.



(a) Approximating a mixture of two Beta distributions using both an exponential and a log-normal component

(b) Approximating a mixture of two Beta distributions using two log-normal components

Figure 4: Comparing the two-component approximations on a Beta mixture distribution.

6 Discussion

References

[KW13] D. P. Kingma and M. Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[Loc+18] F. Locatello et al. "Boosting variational inference: an optimization perspective". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 464–472.

[MFA17] A. C. Miller, N. J. Foti, and R. P. Adams. "Variational boosting: Iteratively refining posterior approximations". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2420–2429.

A Supplementary Plots