Boosting Variational Inference

Fangjian Guo
MIT
guo@csail.mit.edu

Xiangyu Wang
Duke University
xw56@stat.duke.edu

Kai Fan
Duke University
kai.fan@duke.edu

Tamara Broderick
MIT
tbroderick@csail.mit.edu

David Dunson
Duke University
dunson@duke.edu

Abstract

Modern Bayesian inference typically requires some form of posterior approximation, and mean-field variational inference (MFVI) is an increasingly popular choice due to its speed. But MFVI can be inaccurate in various aspects, including an inability to capture multimodality in the posterior and underestimation of the posterior covariance. These issues arise since MFVI considers approximations to the posterior only in a family of factorized distributions. We instead consider a much more flexible approximating family consisting of *all possible finite mixtures* of a parametric base distribution (e.g., Gaussian). In order to efficiently find a high-quality posterior approximation within this family, we borrow ideas from gradient boosting and propose *boosting variational inference* (BVI). BVI iteratively improves the current approximation by mixing it with a new component from the base distribution family. We develop practical algorithms for BVI and demonstrate their performance on both real and simulated data.

1 Introduction

Bayesian inference offers a flexible framework for learning with rich, hierarchical models of data and for coherently quantifying uncertainty in unknown parameters through the posterior distribution. However, for any moderately complex model, the posterior is intractable to calculate exactly and must be approximated. Mean-field variational inference (MFVI) has grown in popularity as a method for approximating the posterior since it is often fast even for large data sets.

MFVI is fast in part because it formulates posterior approximation as an optimization problem, and leads to an efficient coordinate-ascent algorithm when there is certain structure within the model, known as "conditional conjugacy" [3]. Such conjugacy properties typically only hold for factorization approximations, which effectively cannot capture multimodality and underestimate the posterior covariance, sometimes drastically [1, 28, 27, 25, 17]. The linear response technique [13] and the full-rank approach within [15] provide a correction to the covariance underestimation of MFVI in the unimodal case but do not address the multimodality issue. "Black-box" inference, as in [23] and the mean-field approach within [15], focus on making the MFVI optimization problem easier for practitioners, by avoiding tedious calculations, but they do not change the optimization objective of MFVI and therefore still face the problems outlined here.

An alternative and more flexible class of approximating distributions for variational inference (VI) is the family of mixture models. Indeed, even if we consider only Gaussian base distributions, one can find a mixture of Gaussians that is arbitrarily close to *any* continuous probability density [7, 20]. [2, 14, 12] have previously considered using approximating families with a fixed number of mixture components; these authors also employ a further approximation to the VI optimization objective. The resulting optimization algorithms have clear practical limitations, which limit the ability to find a good approximation in the mixture model family. In particular, there is large sensitivity to initial values, for good performance algorithms may need to be rerun for many different initializations and component numbers, and the approximation to the VI objective can limit flexibility.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

As a much more effective algorithm, which automatically adapts the number of mixture components to the complexity of the posterior and increases approximation accuracy, we propose a novel approach inspired by boosting. We call our method *boosting variational inference* (BVI). BVI starts with a single-component approximation and proceeds to add a new mixture component at each step. Independent to our work, we notice that this idea is also considered by a concurrent paper [18].

2 Variational inference and Gaussian mixtures

Suppose we observe N data points, collected in the matrix X with N rows. A Bayesian model is specified through a prior $\pi(\boldsymbol{\theta})$ and likelihood $p(X|\boldsymbol{\theta})$, yielding the posterior $p(\boldsymbol{\theta}|X) \propto \pi(\boldsymbol{\theta})p(X|\boldsymbol{\theta}) =: f(\boldsymbol{\theta})$ by the Bayes Theorem, with $\boldsymbol{\theta} \in \mathbb{R}^D$. While f is easy to evaluate, the normalizing constant $p(X)$ involves an intractable integral, so an approximation is needed. The posterior $p(\boldsymbol{\theta}|X)$ is rarely used directly; rather, we would often like to report a mean, covariance, or other posterior functional $\mathbb{E}_p\, g(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|X)g(\boldsymbol{\theta})d\boldsymbol{\theta}$ for some function g. E.g., $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ yields the posterior mean.

One approach to approximate the posterior is as follows. Choose a discrepancy $\mathcal{D}$ between distribution $q(\boldsymbol{\theta})$ and the exact posterior $p_X(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|X)$. We assume $\mathcal{D}$ is non-negative and zero only when $q = p_X$. In general, though, the optimum $\mathcal{D}^* := \inf_{q\in\mathcal{H}} \mathcal{D}(q, p_X)$ over some constrained family of distributions $\mathcal{H}$ may be strictly greater than zero. Roughly, we expect a larger $\mathcal{H}$ to yield a lower $\mathcal{D}^*$ but at a higher computational cost. When $\mathcal{D}(q, p_X) = \mathcal{D}_{\mathrm{KL}}(q||p_X)$, the Kullback-Leibler (KL) divergence between q and p_X, this optimization problem is called *variational inference*. A particularly flexible choice for $\mathcal{H}$ is the family of *all finite mixtures*. More precisely, let $h_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ be some parametric distribution over $\boldsymbol{\theta}$ with parameter $\boldsymbol{\phi} \in \Phi$. E.g., for a Gaussian mixtures, $h_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ with mean vector $\boldsymbol{\mu}$ and positive semidefinite covariance $\boldsymbol{\Sigma}$. Let Δ_k denote the (k-1)-dimensional simplex: $\Delta_k = \{w \in \mathbb{R}^k : \sum_{j=1}^k w_j = 1 \;\&\; \forall j, w_j \geq 0\}$. Then $\mathcal{H}_k$ is the set of all k-component mixtures over these base distributions, and $\mathcal{H}_\infty$ is the set of *all finite mixtures*, namely

$$\mathcal{H}_k = \{h : h(\boldsymbol{\theta}) = \sum_{j=1}^{k} w_j h_{\phi_j}(\boldsymbol{\theta}), w \in \Delta_k, \boldsymbol{\phi} \in \Phi^k\}, \quad \mathcal{H}_\infty = \bigcup_{k=1}^{\infty} \mathcal{H}_k. \tag{1}$$

Our main contribution is to propose a novel algorithm for approximately solving this discrepancy minimization problem.

3 Boosting and Gradient Boosting

In general, reaching $\mathcal{D}^*$ may require an infinite mixture. We consider a greedy, incremental procedure, as in [29], to approach $\mathcal{D}^*$ with a sequence of finite mixtures $q_1, q_2, \cdots$ for each $q_t \in \mathcal{H}_t$. The quality of approximation can be measured with the excess discrepancy $\Delta\mathcal{D}(q_t) := \mathcal{D}(q_t, p_X) - \mathcal{D}^* \geq 0$, and we would like $\Delta\mathcal{D}(q_t) \to 0$. Hence, given any $\epsilon > 0$, we can find a large enough t such that $\Delta\mathcal{D}(q_t) \leq \epsilon$. In particular, we start with a single base distribution $q_1 = h_{\phi_1}$ for some $\boldsymbol{\phi}_1$. Iteratively, at each step $t = 2, 3, \ldots$, let q_{t-1} be the approximation from the previous step. Form q_t by mixing a new base distribution h_t with weight $\alpha_t \in [0, 1]$ together with q_{t-1} with weight $(1 - \alpha_t)$. This approach is called *greedy* [22, Ch. 4] if we choose (approximately) optimal base distribution h_t and weight α_t at each step t:

$$q_t = (1 - \alpha_t)q_{t-1} + \alpha_t h_t, \quad \mathcal{D}(q_t, p_X) \leq \inf_{\phi\in\Phi, 0\leq\alpha\leq 1} \mathcal{D}((1 - \alpha)q_{t-1} + \alpha h_\phi, p_X) + \epsilon_t, \tag{2}$$

where we relax optimality to within some non-negative sequence $\epsilon_t \searrow 0$. At each step q_t remains normalized by construction and takes the form of a mixture of base distributions. The iterative updates are in the style of *boosting* or *greedy error minimization* [9, 8, 10, 16]. Under convexity and strong smoothness conditions on $\mathcal{D}(\cdot, p_x)$, *Theorem II.1* of [29] guarantees that $\Delta\mathcal{D}(q_t)$ converges to zero at rate $O(1/t)$. We will verify that KL divergence satisfies these conditions in Theorem 1.

Gradient Boosting Let $\mathcal{D}(q) := \mathcal{D}(q, p_X)$ as a shorthand notation. Rather than jointly optimizing $\mathcal{D}((1-\alpha_t)q_{t-1}+\alpha_t h_t)$ over (α_t, h_t), which may be non-convex and difficult in general, we consider nearly optimal choices (cf. Eq. (2)). We choose h_t first, in a gradient descent style, and then optimize the corresponding weight α_t. For h_t, we follow gradient boosting [11] and consider the *functional gradient* $\nabla\mathcal{D}(q)$ at the current solution $q = q_{t-1}$. In what follows, we adopt the notation $\langle g, h\rangle = \int g(\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\|h\|_2^2 = \int h(\boldsymbol{\theta})^2 d\boldsymbol{\theta}$. When $\|h\|_2 \approx 0$, a Taylor expansion yields

$$\mathcal{D}(q + h) = \mathcal{D}(q) + \langle g, h\rangle + o(\|h\|_2^2), \tag{3}$$

where g is the functional gradient $\nabla\mathcal{D}(q) := g(\boldsymbol{\theta})$. We choose h to minimize the inner product in Eq. (3); i.e., as in gradient descent, we choose h to "match the direction" of $-\nabla\mathcal{D}(q)$.

4 Boosting Variational Inference

Boosting variational inference (BVI) applies the framework of the previous section with Kullback-Leibler (KL) divergence as the discrepancy measure. We first justify the choice of KL, and then present a two-stage multivariate Gaussian mixture boosting algorithm (Algorithm 1) to stochastically decrease the excess discrepancy. In each iteration, it firstly identifies h_t with gradient boosting, and then solves for α_t. The KL discrepancy measure is defined as

$$\mathcal{D}(q, p_X) = \mathcal{D}_{\mathrm{KL}}(q\|p_X) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p_X(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \log p(X) + \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}. \tag{4}$$

By dropping the constant $\log p(X)$, an effective discrepancy (negative value of ELBO [3]) can be defined as $\tilde{\mathcal{D}}_{\mathrm{KL}}(q) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$. The following theorem shows that KL satisfies the greedy boosting conditions [29, 22] — under some additional assumptions (that may, e.g., hold on a bounded set). See Appendix A for proof and discussions. This trivially implies that the conditions also hold for $\tilde{\mathcal{D}}_{\mathrm{KL}}(q)$. See Appendix C for a similar analysis of KL divergence in the other direction, $\mathcal{D}(p_X\|q)$.

Theorem 1. *Given densities q_1, q_2 and true density p, KL divergence is a convex functional, i.e., for any $\alpha \in [0, 1]$ satisfying*

$$\mathcal{D}_{KL}((1-\alpha)q_1 + \alpha q_2\|p) \leq (1-\alpha)\,\mathcal{D}_{KL}(q_1\|p) + \alpha\,\mathcal{D}_{KL}(q_2\|p). \tag{5}$$

If we further assume that densities are bounded $q_1(\boldsymbol{\theta}), q_2(\boldsymbol{\theta}) \geq a > 0$, and denote the functional gradient of KL at density q as $\nabla\,\mathcal{D}_{KL}(q) = \log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})$, then the KL divergence is also strongly smooth, i.e., satisfying

$$\mathcal{D}_{KL}(q_2\|p) - \mathcal{D}_{KL}(q_1\|p) \leq \langle \nabla\,\mathcal{D}_{KL}(q_1\|p), q_2 - q_1\rangle + \frac{1}{a}\|q_2 - q_1\|_2^2. \tag{6}$$

Setting α_t with Stochastic Newton For fixed h_t, $\tilde{\mathcal{D}}_{\mathrm{KL}}$ is a *convex* function of α_t (see (11) in Appendix B). We can estimate both its first and second derivatives with Monte Carlo, by drawing samples from h_t and q_{t-1}. Then we can use a stochastic 2^{nd}-order method, e.g., Newton's [5], to solve α_t. See Appendix B for details.

Setting h_t with Laplacian Gradient Boosting Following the idea of Eq. (3) for gradient boosting, we take the Taylor expansion of $\tilde{\mathcal{D}}_{\mathrm{KL}}(q_t)$ around $\alpha_t \searrow 0$

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(q_t) = \tilde{\mathcal{D}}_{\mathrm{KL}}(q_{t-1}) + \alpha_t\langle h_t, \log\frac{q_{t-1}}{f}\rangle - \alpha_t\langle q_{t-1}, \log\frac{q_{t-1}}{f}\rangle + o(\alpha_t^2), \tag{7}$$

which suggests minimizing $\langle h_t, \log\frac{q_{t-1}}{f}\rangle$, where $\log\frac{q_{t-1}}{f}$ is the functional gradient $\nabla\,\tilde{\mathcal{D}}_{\mathrm{KL}}(q_{t-1})$.

However, direct minimization of the inner product is ill-posed since h_t will degenerate to a point mass at the minimum of functional gradient. Instead, following the least square procedure of [11], we "match the direction" in terms of l_2 norm by $\hat{h}_t = \arg\min_{h=h_\phi,\lambda>0} \|\lambda h - \log(f/q_{t-1})\|_2^2$. Plugging in the optimal value for λ, and with some algebra the objective is identical to

$$\hat{h}_t = \arg\max_{h=h_\phi} \mathbb{E}_{\boldsymbol{\theta}\sim h} \log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) - 1/2 \cdot \log\|h\|_2^2, \tag{8}$$

where $\log(f/q_{t-1})$ is effectively the *residual log-likelihood*. Ideally, when $q_{t-1} \propto f$, the residual should be flat; otherwise, its has peaks where posterior density is underestimated and basins where overestimated. For Gaussian family $h_\phi(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, (8) becomes

$$\hat{h}_{\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} \log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) + 1/4 \cdot \log|\boldsymbol{\Sigma}|. \tag{9}$$

We notice that the log determinant term prevents degeneration. We therefore propose the following heuristic (Algorithm 1) to efficiently optimize (9): we approximately decompose the residual $\log(f/q_{t-1})$ into a constant plus a quadratic peak $-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\eta})^T\boldsymbol{S}^{-1}(\boldsymbol{\theta}-\boldsymbol{\eta})$ (i.e. *Laplacian approximation* to f/q_{t-1}), and then (9) becomes convex and solutions are in closed-form as $\boldsymbol{\mu}^* = \boldsymbol{\eta}$, $\boldsymbol{\Sigma}^* = \boldsymbol{S}/2$, where $\boldsymbol{\eta}$ and $\boldsymbol{S}^{-1}$ can be solved numerically with any suitable optimization routine.

Algorithm 1 Laplacian Gradient Boosting for Gaussian Base Family $\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$

Require: evaluable product density of prior and likelihood $f(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathbb{R}^D$

Start with some initial approximation $q_1 = \mathcal{N}_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1}$ ▷ e.g. $q_1 = \mathcal{N}_{\mathbf{0},c\boldsymbol{I}}$

for $t = 2$ to T **do**

$\hat{\boldsymbol{\mu}}_t \leftarrow \arg\min_{\boldsymbol{\theta}} \log(q_{t-1}(\boldsymbol{\theta})/f(\boldsymbol{\theta}))$ with an optimization routine initialized at $\boldsymbol{\theta}_0 \sim q_{t-1}$.

$\boldsymbol{H}_t \leftarrow \text{Hessian}_{\boldsymbol{\theta}=\hat{\boldsymbol{\mu}}_t} \log(q_{t-1}(\boldsymbol{\theta})/f(\boldsymbol{\theta}))$ with numerical approximation

$\hat{\boldsymbol{\Sigma}}_t \leftarrow \boldsymbol{H}_t^{-1}/2$ and let $\hat{h}_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\Sigma}}_t)$ be the new component

$\hat{\alpha}_t \leftarrow \arg\min_{\alpha_t} \tilde{\mathcal{D}}_{\text{KL}}((1-\alpha_t)q_{t-1} + \alpha_t \hat{h}_t)$ with Algorithm 2. ▷ See Appendix B

$q_t \leftarrow (1-\hat{\alpha}_t)q_{t-1} + \hat{\alpha}_t \hat{h}_t$. ▷ Boosting

end for

return q_t

5 Experiments

In this section, we compare performance of BVI to MFVI with both toy and real-world data.

Toy Examples Figure 1 highlights the ability of BVI (unlike MFVI) to capture (a) heavy tails (b) multimodality and (c) multivariate distributions. Figure 1 (a) is a Cauchy density $p_X(\theta) \propto \frac{1}{1+(\theta/2)^2}$; (b) is a mixture of univariate Gaussians with different locations and scales; (c) is a mixture of five 2D-Gaussians with random locations and covariances. We initialize with a Gaussian with very large (co)variance, and then run BVI for 50 iterations. Sequences $(\alpha_t)_t$ and $(\hat{\mathcal{D}}_{\text{KL}}(q_t, p))_t$ are shown in subplots. Since these distributions are non-conjugate, we run the *automatic variational inference* (ADVI) in *Stan* [15, 6] to obtain results for MFVI (orange).

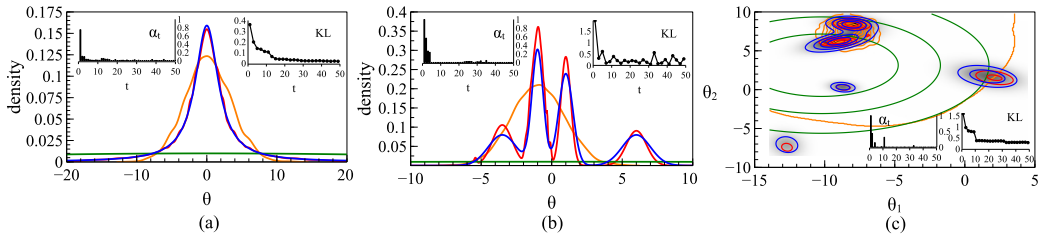

Figure 1: Toy examples: (a) (heavy-tailed) Cauchy distribution (b) a mixture of four univariate Gaussians and (c) a mixture of five bivariate Gaussians with random means and covariances. Curves/contours are colored by blue (true), red (BVI), green (initial q_1 in BVI) and orange (ADVI). Sequence of $(\alpha_t)_t$ and Monte Carlo estimates of $(\mathcal{D}_{\text{KL}}(q_t))_t$ are plotted against iteration in subplots.

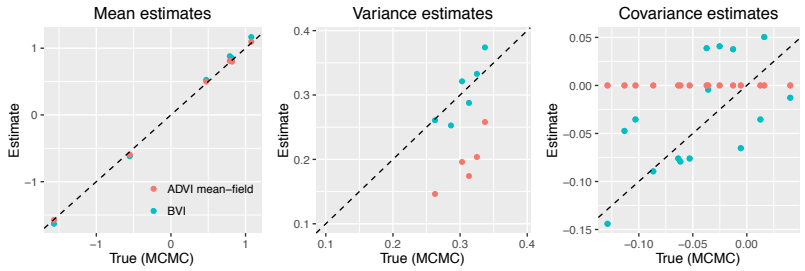

Figure 2: Estimated posterior mean and covariance for logistic regression (dashed: estimate = true).

Bayesian Logistic Regression We apply our algorithm to Bayesian logistic regression for the `Nodal` dataset [4], consisting of $N = 53$ observations of six predictors $\boldsymbol{x}_i$ (intercept included) and a binary response $y_i \in \{-1, +1\}$. The likelihood is $\prod_{i=1}^N g(y_i \mathbf{x}_i^\top \boldsymbol{\beta})$, where $g(x) = (1+e^{-x})^{-1}$ and we use the prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. For reference, we show results from the Polya-Gamma sampler (an MCMC algorithm for logistic regression) using R package `BayesLogit` [21] as the ground truth. We also compare the performance of BVI to MFVI (ADVI with *Stan*). As shown in Figure 2, while both methods capture the correct mean, BVI provides better estimates of the variance and, unlike MFVI, does not set the covariances to zero. We expect more dramatic differences in cases where MFVI yields biased estimates of the posterior means [26] and cases where the posterior is multimodal. We plan to investigate these cases in future work.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 10. Springer, New York, 2006.

[2] C. M. Bishop, N. Lawrence, T. Jaakkola, and M. I. Jordan. Approximating posterior distributions in belief networks using mixtures. In *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, volume 10, page 416, 1998.

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

[4] B. Brown. Prediction analyses for binary data. *Biostatistics Casebook*, pages 3–18, 1980.

[5] R. H. Byrd, S. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[6] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *J Stat Softw*, 2016.

[7] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[8] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[9] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[10] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[11] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[12] S. Gershman, M. Hoffman, and D. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning*, pages 663–670, New York, NY, USA, 2012.

[13] R. J. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.

[14] T. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pages 163–173. Springer, 1998.

[15] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Automatic variational inference in stan. In *Advances in neural information processing systems*, pages 568–576, 2015.

[16] J. Q. Li and A. R. Barron. Mixture density estimation. In *Advances in neural information processing systems*, pages 279–285, 1999.

[17] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter 33. Cambridge University Press, 2003.

[18] A. C. Miller, N. Foti, and R. P. Adams. Variational boosting: Iteratively refining posterior approximations. *arXiv preprint arXiv:1611.06585*, 2016.

[19] T. Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[20] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[21] N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[22] A. Rakhlin. *Applications of empirical processes in learning theory: algorithmic stability and generalization bounds*. PhD thesis, Massachusetts Institute of Technology, 2006.

[23] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[24] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[25] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 71(2):319–392, 2009.

[26] R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Workshop on Inference and Estimation in Probabilistic Time-Series Models*, volume 2, 2008.

[27] R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, A. T. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*. Cambridge University Press, 2011.

[28] B. Wang and M. Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380, 2004.

[29] T. Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

Appendices

A Proof of Theorem 1

Proof. For any densities q_1, q_2 and $\alpha \in [0, 1]$, we have the convexity from

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}((1-\alpha)q_1 + \alpha q_2 \| p) &= (1-\alpha)\int q_1 \log \frac{(1-\alpha)q_1 + \alpha q_2}{p} \mathrm{d}\boldsymbol{\theta} + \alpha \int q_2 \log \frac{(1-\alpha)q_1 + \alpha q_2}{p} \mathrm{d}\boldsymbol{\theta} \\
&= (1-\alpha)\int q_1 \log \frac{q_1[1+\alpha(q_2/q_1 - 1)]}{p} \mathrm{d}\boldsymbol{\theta} + \alpha \int q_2 \log \frac{q_2[1+(1-\alpha)(q_1/q_2 - 1)]}{p} \mathrm{d}\boldsymbol{\theta} \\
&= (1-\alpha)\int q_1 \log \frac{q_1}{p} \mathrm{d}\boldsymbol{\theta} + \alpha \int q_2 \log \frac{q_2}{p} \mathrm{d}\boldsymbol{\theta} + (1-\alpha)\int q_1 \log[1+\alpha(q_2/q_1 - 1)] \mathrm{d}\boldsymbol{\theta} \\
&\quad + \alpha \int q_2 \log[1+(1-\alpha)(q_1/q_2 - 1)] \mathrm{d}\boldsymbol{\theta} \\
&\leq (1-\alpha)\,\mathcal{D}_{\mathrm{KL}}(q_1 \| p) + \alpha\,\mathcal{D}_{\mathrm{KL}}(q_2 \| p) + (1-\alpha)\int \alpha(q_2 - q_1) \mathrm{d}\boldsymbol{\theta} + \alpha \int (1-\alpha)(q_1 - q_2) \mathrm{d}\boldsymbol{\theta} \\
&\leq (1-\alpha)\,\mathcal{D}_{\mathrm{KL}}(q_1 \| p) + \alpha\,\mathcal{D}_{\mathrm{KL}}(q_2 \| p),
\end{aligned}$$

where we have used $\log(1+x) \leq x$ and $\int (q_2 - q_1) \mathrm{d}\boldsymbol{\theta} = 0$.

Let $h = q_2 - q_1$, then we have $\int h(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int (q_2 - q_1) \mathrm{d}\boldsymbol{\theta} = 0$. Again, using the inequality $\log(1+x) \leq x$, we have the strong smoothness from

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(q_1 + h \| p) - \mathcal{D}_{\mathrm{KL}}(q_1 \| p) &= \langle q_1 + h, \log(q_1 + h) \rangle - \langle q_1, \log q_1 \rangle - \langle h, \log p \rangle \\
&\leq \langle q_1 + h, \frac{h}{q_1} + \log q_1 \rangle - \langle q_1, \log q_1 \rangle - \langle h, \log p \rangle \\
&= \langle h, \log q_1 \rangle + \langle h, \frac{h}{q_1} \rangle - \langle h, \log p \rangle \\
&\leq \langle \log q_1 - \log p, h \rangle + \frac{1}{a} \|h\|_2^2,
\end{aligned}$$

where in the last step we used the assumption that $q_1, q_2 \geq a > 0$. □

Note that here we assume that densities are lower-bounded by $a > 0$. While this may seem unrealistic for some densities, this is a technical requirement to ensure that the discrepancy is "smooth" enough so that greedy boosting can be applied. For densities on $\mathbb{R}^D$ that do not have a lower bound, we suggest consider approximation within $(1-\epsilon)$ of probability mass. In particular, consider a closed set $\boldsymbol{\Theta} \subset \mathbb{R}^D$ such that $\int_{\boldsymbol{\Theta}} p_x(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = (1-\epsilon)$ for a small $\epsilon > 0$. Then we can use the smallest density in $\boldsymbol{\Theta}$ as a lower bound.

B Stochastic Newton's Algorithm for Solving Mixing Weights

Fixing h_t, by taking derivatives of effective discrepancy with respect to α_t, we have

$$\frac{\partial\, \tilde{\mathcal{D}}_{\mathrm{KL}}(q_t)}{\partial \alpha_t} = \int (h_t - q_{t-1}) \log \frac{(1-\alpha_t)q_{t-1} + \alpha_t h_t}{f} \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim h_t} \gamma_{\alpha_t}(\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim q_{t-1}} \gamma_{\alpha_t}(\boldsymbol{\theta}), \quad (10)$$

$$\frac{\partial^2\, \tilde{\mathcal{D}}_{\mathrm{KL}}(q_t)}{\partial \alpha_t^2} = \int \frac{(h_t - q_{t-1})^2}{(1-\alpha_t)q_{t-1} + \alpha_t h_t} \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim h_t} \eta_{\alpha_t}(\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim q_{t-1}} \eta_{\alpha_t}(\boldsymbol{\theta}) \geq 0, \quad (11)$$

where γ_{α_t} and η_{α_t} are evaluable functions of $\boldsymbol{\theta}$ and α_t given below in Algorithm 2. Because the first and second derivatives are stochastically estimated instead of exact, to ensure convergence we use a decaying sequence of step sizes b/k in Algorithm 2 that satisfy the Robbins-Monro conditions [5, 24].

Algorithm 2 Stochastic Newton's

Require: current approximation $q_{t-1}(\boldsymbol{\theta})$, new component $h_t(\boldsymbol{\theta})$, product density of prior and likelihood $f(\boldsymbol{\theta})$, Monte Carlo sample size n, initial step size $b > 0$
Initialize $k \leftarrow 0$, $\alpha_t \leftarrow 0$
Independently draw $\{\boldsymbol{\theta}_i^{(h)}\} \sim h_t$ and $\{\boldsymbol{\theta}_i^{(q)}\} \sim q_{t-1}$ for $i = 1, \cdots, n$
while α_t not convergent **do**
  Compute $\hat{\mathcal{D}}'_{\mathrm{KL}} = \frac{1}{n}\sum_i(\gamma_{\alpha_t}(\boldsymbol{\theta}_i^{(h)}) - \gamma_{\alpha_t}(\boldsymbol{\theta}_i^{(q)}))$ with $\gamma_{\alpha_t}(\boldsymbol{\theta}) = \log\frac{(1-\alpha_t)q_{t-1}(\boldsymbol{\theta})+\alpha_t h_t(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}$
  Compute $\hat{\mathcal{D}}''_{\mathrm{KL}} = \frac{1}{n}\sum_i(\eta\alpha_t(\boldsymbol{\theta}_i^{(h)}) - \eta\alpha_t(\boldsymbol{\theta}_i^{(q)}))$ with $\eta_{\alpha_t}(\boldsymbol{\theta}) = \frac{h_t(\boldsymbol{\theta})-q_{t-1}(\boldsymbol{\theta})}{(1-\alpha_t)q_{t-1}(\boldsymbol{\theta})+\alpha_t h_t(\boldsymbol{\theta})}$
  $k \leftarrow k+1$, $\alpha_t \leftarrow \alpha_t - (b/k)\hat{\mathcal{D}}'_{\mathrm{KL}}/\hat{\mathcal{D}}''_{\mathrm{KL}}$
end while

C BVI with KL Divergence in the Alternative Direction

KL divergence is asymmetric and can be written in two directions [19]: $\mathcal{D}_{\mathrm{KL}}(q_t\|p_X)$ and $\mathcal{D}_{\mathrm{KL}}(p_X\|q_t)$. We have discussed BVI algorithm in the main text based on the "exclusive" direction $\mathcal{D}_{\mathrm{KL}}(q_t\|p_X)$; in this Appendix, we show that a similar algorithm can be derived with the alternative direction, namely the "inclusive" direction

$$\mathcal{D}_{\mathrm{KL}}(p_X\|q_t) = \int p_X(\boldsymbol{\theta})\log\frac{p_X(\boldsymbol{\theta})}{q_t(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} = \int p_X(\boldsymbol{\theta})\log p_X(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - c\int f(\boldsymbol{\theta})\log q_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \quad (12)$$

Here $c > 0$ is the normalizing constant. Hence, by dropping constants that do not involve q_t, we can similarly define an effective discrepancy as

$$\tilde{\mathcal{D}}_{kl}(q) = -\int f(\boldsymbol{\theta})\log q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \quad (13)$$

where we use lower-case italic subscript *kl* to distinguish it from the previous KL divergence. Fixing p_X, we will use the notation $\mathcal{D}_{kl}(q) := \mathcal{D}_{\mathrm{KL}}(p_X\|q)$.

Before proceeding to the derivation of algorithm, we firstly show that $\mathcal{D}_{\mathrm{KL}}(p_X\|q_t)$ also satisfies the regularity conditions required by greedy boosting.

Lemma 1. *$\mathcal{D}_{KL}(p\|q)$ is a convex functional of q, i.e., given densities p, q_1, q_2, it holds that*

$$\mathcal{D}_{KL}(p\|(1-\alpha)q_1 + \alpha q_2) \le (1-\alpha)\,\mathcal{D}_{KL}(p\|q_1) + \alpha\,\mathcal{D}_{KL}(p\|q_2) \quad (14)$$

for all $\alpha \in [0,1]$.

Proof. For any $\alpha \in [0,1]$, by Jensen's inequality,

$$\log[(1-\alpha)q_1 + \alpha q_2] \ge (1-\alpha)\log q_1 + \alpha\log q_2.$$

Then it directly follows that

$$\mathcal{D}_{\mathrm{KL}}(p\|(1-\alpha)q_1 + \alpha q_2) \le (1-\alpha)\,\mathcal{D}_{\mathrm{KL}}(p\|q_1) + \alpha\,\mathcal{D}_{\mathrm{KL}}(p\|q_2).$$

□

Another condition required by boosting framework of [29] is strong smoothness. It has been shown that for $\mathcal{D}_{\mathrm{KL}}(p\|q)$ to satisfy strong smoothness, it suffices to require an upper bound on the log ratio of base family densities [16] [22, Chapter 4]:

$$a = \sup_{\substack{q_1,q_2\in\{h_\phi:\phi\in\Phi\},\\ \boldsymbol{\theta}\in\mathbb{R}^D}} \log\frac{q_1(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})} < +\infty. \quad (15)$$

This condition might be translated to constraints on the parameter space Φ.

Solving α_t Again, we firstly show that by fixing h_t, optimizing α_t is a convex problem, and can be solved with stochastic Newton's method. Since in $\tilde{\mathcal{D}}_{kl}$ the integral measure does not involve α_t, we have

$$\frac{\partial}{\partial\alpha}\tilde{\mathcal{D}}_{kl} = -\int f(\boldsymbol{\theta})\frac{\partial \log q_t(\boldsymbol{\theta})}{\partial\alpha}d\boldsymbol{\theta} = -\int f(\boldsymbol{\theta})\frac{h_t - q_{t-1}}{(1-\alpha)q_{t-1}+\alpha h_t}\mathrm{d}\boldsymbol{\theta}, \tag{16}$$

$$\frac{\partial^2}{\partial\alpha^2}\tilde{\mathcal{D}}_{kl} = -\int f(\boldsymbol{\theta})\frac{\partial^2 \log q_t(\boldsymbol{\theta})}{\partial\alpha^2}d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\frac{(h_t - q_{t-1})^2}{[(1-\alpha)q_{t-1}+\alpha h_t]^2}\mathrm{d}\boldsymbol{\theta} \geq 0. \tag{17}$$

Notice that as we cannot sample from the true posterior $p_X(\boldsymbol{\theta}) \propto f(\boldsymbol{\theta})$, we cannot directly estimate these derivatives with Monte Carlo. However, importance sampling can be applied here, by rewriting the derivatives as

$$\frac{\partial}{\partial\alpha}\tilde{\mathcal{D}}_{kl} = -\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim q_{t-1}}\frac{f(\boldsymbol{\theta})}{q_{t-1}(\boldsymbol{\theta})}\frac{h_t(\boldsymbol{\theta}) - q_{t-1}(\boldsymbol{\theta})}{(1-\alpha)q_{t-1}(\boldsymbol{\theta})+\alpha h_t(\boldsymbol{\theta})} \tag{18}$$

$$\frac{\partial^2}{\partial\alpha^2}\tilde{\mathcal{D}}_{kl} = \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim q_{t-1}}\frac{f(\boldsymbol{\theta})}{q_{t-1}(\boldsymbol{\theta})}\frac{(h_t(\boldsymbol{\theta}) - q_{t-1}(\boldsymbol{\theta}))^2}{[(1-\alpha)q_{t-1}(\boldsymbol{\theta})+\alpha h_t(\boldsymbol{\theta})]^2}. \tag{19}$$

Hence we can draw "particles" $\boldsymbol{\theta}_i \sim q_{t-1}(\boldsymbol{\theta})$ and reweigh it with $w_i = f(\boldsymbol{\theta}_i)/q_{t-1}(\boldsymbol{\theta}_i)$. And we have the following algorithm.

Algorithm 3 Stochastic Newton's with $\mathcal{D}_{kl}(q_t) = \mathcal{D}_{\mathrm{KL}}(p_x\|q_t)$

Require: current approximation $q_{t-1}(\boldsymbol{\theta})$, new component $h_t(\boldsymbol{\theta})$, product density of prior and likelihood $f(\boldsymbol{\theta})$, importance sampling sample size n, initial step size $b > 0$
  Initialize $k \leftarrow 0$, $\alpha_t \leftarrow 0$
  Independently draw $\{\boldsymbol{\theta}_i\} \sim q_{t-1}$ for $i = 1, \cdots, n$
  $w_i \leftarrow f(\boldsymbol{\theta}_i)/q_{t-1}(\boldsymbol{\theta}_i)$ for $i = 1, \cdots, n$
  while α_t not convergent **do**
    $s_i \leftarrow \frac{h_t(\boldsymbol{\theta}_i) - q_{t-1}(\boldsymbol{\theta}_i)}{(1-\alpha_t)q_{t-1}(\boldsymbol{\theta}_i)+\alpha_t h_t(\boldsymbol{\theta}_i)}$ for $i = 1, \cdots, n$
    Compute $\hat{\mathcal{D}}'_{kl} = -\frac{1}{n}\sum_i w_i s_i$
    Compute $\hat{\mathcal{D}}''_{kl} = \frac{1}{n}\sum_i w_i s_i^2$
    $k \leftarrow k+1$, $\alpha_t \leftarrow \alpha_t - (b/k)\hat{\mathcal{D}}'_{kl}/\hat{\mathcal{D}}''_{kl}$
  end while

Setting h_t It turns out with $\mathcal{D}_{kl}$ we can derive the *same* gradient boosting procedure as Algorithm 1. Again, with Taylor expansion of the effective discrepancy at $\alpha_t \searrow 0$, we have

$$\tilde{\mathcal{D}}_{kl}(q_t) = \tilde{D}_{kl}(q_{t-1}) - \alpha_t\langle f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta}), h_t(\boldsymbol{\theta})\rangle + \alpha_t\int f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + o(\alpha_t^2). \tag{20}$$

Similarly, to avoid degeneracy, we consider

$$\min_{h=h_\phi,\lambda>0}\|\lambda h - f/q_{t-1}\|_2^2, \tag{21}$$

which is equivalent to

$$\max_{h=h_\phi}\log\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim h}(L(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) - 1/2\cdot\log\|h\|_2^2. \tag{22}$$

By Jensen's inequality, we observe that the previous objective in Eqs. (8) is in fact a *lower bound* of this objective, namely

$$\log\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim h}(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) - 1/2\cdot\log\|h\|_2^2 \geq \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim h}\log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) - 1/2\cdot\log\|h\|_2^2. \tag{23}$$

And we further notice that this bound is *iteratively tightened*, since as $q_{t-1}(\boldsymbol{\theta})$ better approximates $p_X(\boldsymbol{\theta})$, $f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta}) \propto p_X(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})$ will converge to a constant. By maximizing the lower bound, we arrive at the same gradient boosting objective, and hence the same Laplacian gradient boosting subroutine for setting the new component h_t.

To summarize, with KL divergence in the alternative direction $\mathcal{D}_{\mathrm{KL}}(p_X\|q_t)$ as the discrepancy measure, we derive an algorithm almost identical to Algorithm 1, except that the step for solving $\hat{\alpha}_t$ is performed with Algorithm 3.