Boosting Variational Inference

Richard (Fangjian) Guo*, Xiangyu Wang†, Kai Fan†,
Tamara Broderick*, David B. Dunson†

*Massachusetts Institute of Technology, †Duke University

arXiv:1611.05559

Overview

Black-box Bayesian inference is hard:

- MCMC can be slow.
- Variational inference can be inaccurate.

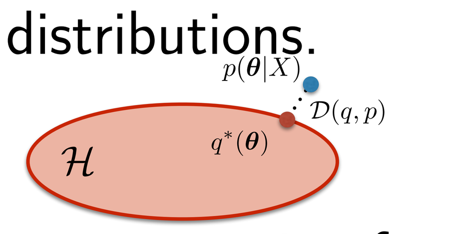
Boosting Variational Inference:

- **Fast**: optimization-based
- **Nonparametric & Adaptive**: iteratively improves by **adapting to residual**

Variational Bayes

Variational Bayes approximates true posterior $p(\boldsymbol{\theta}|X)$ within the closest $q(\boldsymbol{\theta})$ within a family of distributions $\mathcal{H}$, in terms of discrepancy measure $\mathcal{D}$ between the two distributions.

$$q^* = \arg\min_{q \in \mathcal{H}} \mathcal{D}(q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|X))$$



Kullback-Leibler (KL) divergence is often used as discrepancy measure ($f = \pi(\boldsymbol{\theta})p(X|\boldsymbol{\theta})$)

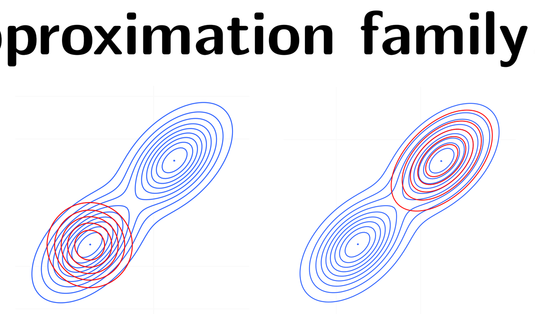
$$\mathcal{D}(q,p) := \mathcal{D}_{\mathsf{KL}}(q\|p) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|X)} \mathrm{d}\boldsymbol{\theta}$$
$$= \text{const} + \int q \log(q/f) \mathrm{d}\boldsymbol{\theta}.$$

Limitations of current VB:

- Point estimates: often good, can be biased
- Poor uncertainty estimates: covariance, multimodality
- Cannot improve accuracy given more time

VB Approximation Family

Accuracy of VB is *mainly* limited by the **inflexibility of approximation family**.



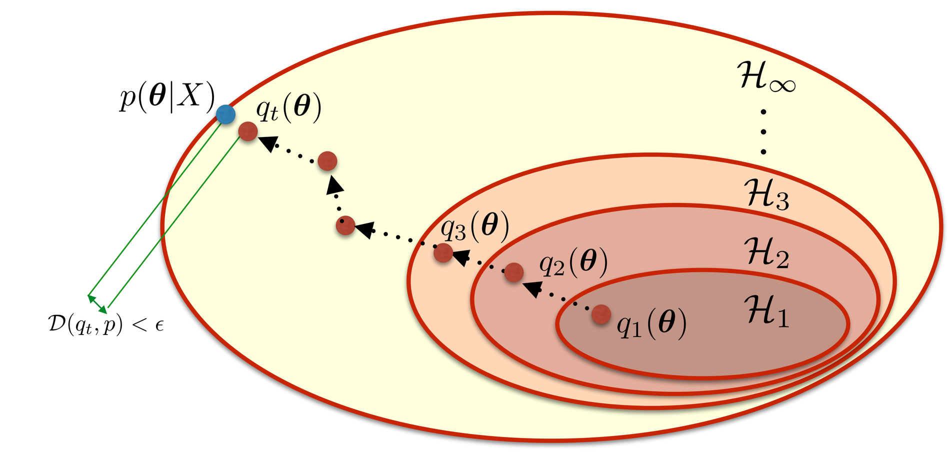
1. Mean-field $q(\boldsymbol{\theta}) = \Pi_i q_i(\theta_i)$
2. Full-rank Gaussian
$$\mathcal{H}_1 = \{h : h(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})\}$$
3. Mixture of k Gaussians
$$\mathcal{H}_k = \{h : h(\boldsymbol{\theta}) = \sum_{j=1}^{k} w_j \mathcal{N}_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{\theta}), \boldsymbol{w} \in \Delta_k\}$$
4. Our choice: All finite Gaussian mixtures
$$\mathcal{H}_\infty = \bigcup_{k=1}^{\infty} \mathcal{H}_k$$

Family	Covariance	Multimodality	Arbitrary approximation
Mean-field	✗	✗	✗
Full-rank $\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$	✓	✗	✗
$\mathcal{H}_k$	✓	✓	✗
$\mathcal{H}_\infty$	✓	✓	✓

Greedy Boosting

Want to construct a sequence of approximations $q_t(\boldsymbol{\theta}) \in \mathcal{H}_t$ such that as $t \to \infty$

$$\Delta\mathcal{D}(q_t) := \mathcal{D}(q_t,p) - \inf_{q \in \mathcal{H}_\infty} \mathcal{D}(q,p) \searrow 0.$$



Greedy Boosting Algorithm

1. Start with $q_1 \in \mathcal{H}_1$.
2. Then iteratively for $t = 2, 3, \cdots$, we let
$$\boxed{q_t = (1-\alpha_t)\, q_{t-1} + \alpha_t\, h_t}$$
such that for some $\epsilon_t \searrow 0$,
$$\mathcal{D}(q_t,p) \leq \inf_{h \in \mathcal{H}_1, 0 \leq \alpha \leq 1} \mathcal{D}((1-\alpha)q_{t-1} + \alpha h, p) + \epsilon_t. \; (*)$$

However, optimization $(*)$ is non-convex.

Our Algorithm

Two-step approach for *Greedy Boosting* $(*)$.

Step 1: Gradient Boosting: Dist. h_t

Friedman, (2001) proposed identifying the form of h_t with the **gradient information** when **increment is small**.

For $\mathcal{D}_{\mathsf{KL}}$, the **negative functional gradient** is the **residual** of log posterior density:

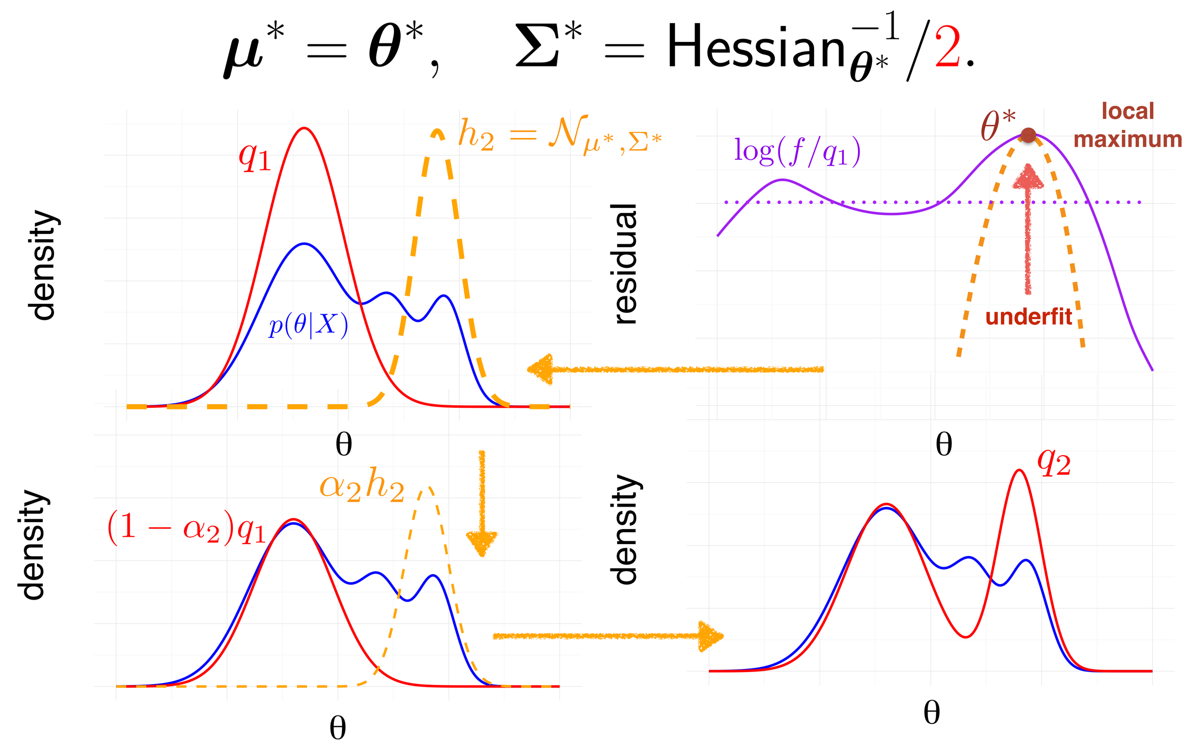
$$-\nabla\, \tilde{\mathcal{D}}_{\mathsf{KL}}(q_{t-1}) = \log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})).$$

To minimize KL, we match h_t to $-\nabla\, \tilde{\mathcal{D}}_{\mathsf{KL}}(q_{t-1})$:

$$\hat{h}_t = \arg\min_{h \in \mathcal{H}_1,\; c>0} \|c \cdot h - \log(f/q_{t-1})\|_2^2.$$

With Laplacian approximation to the residual, we have a simple algorithm for quickly identifying $\hat{h}_{\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t}$ using optimization. We have closed-form solutions:

$$\boldsymbol{\mu}^* = \boldsymbol{\theta}^*, \quad \boldsymbol{\Sigma}^* = \text{Hessian}_{\boldsymbol{\theta}^*}^{-1}/2.$$



Step 2: Stochastic Newton's: Weight α_t

Fixing h_t, determining corresponding weight

$$\alpha_t = \min_{0 \leq \alpha \leq 1} \tilde{\mathcal{D}}_{\mathsf{KL}}((1-\alpha)q_{t-1} + \alpha h_t)$$

is **convex**. Further, by drawing samples from q_{t-1} and h_t, we can get **Monte Carlo estimates** of derivatives $\hat{\mathcal{D}}'_{\mathsf{KL}}$ and $\hat{\mathcal{D}}''_{\mathsf{KL}}$.

Theoretical Results

From Zhang, (2003), for greedy boosting, if $\mathcal{D}(q,p)$ is (1) **convex** in q and (2) **strongly smooth** in q, then we have

$$\Delta\mathcal{D}(q_t) \to 0 \text{ at rate } O(1/t).$$

In **Theorem** 1, we showed that under mild conditions (e.g., that hold on a bounded set) $\mathcal{D}_{\mathsf{KL}}$ **satisfies these conditions**.

Simulation Experiments

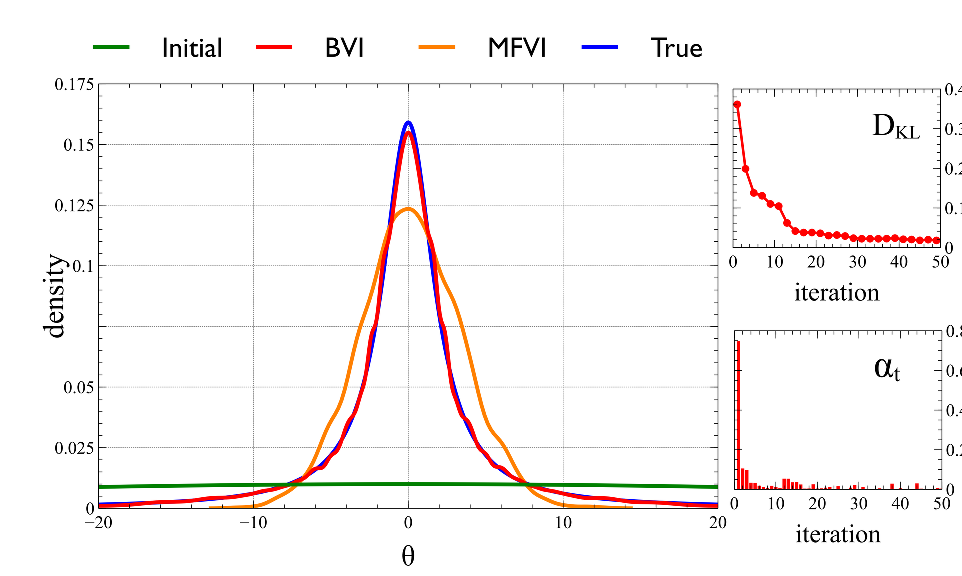


Figure 1: True: Heavy-tailed Cauchy

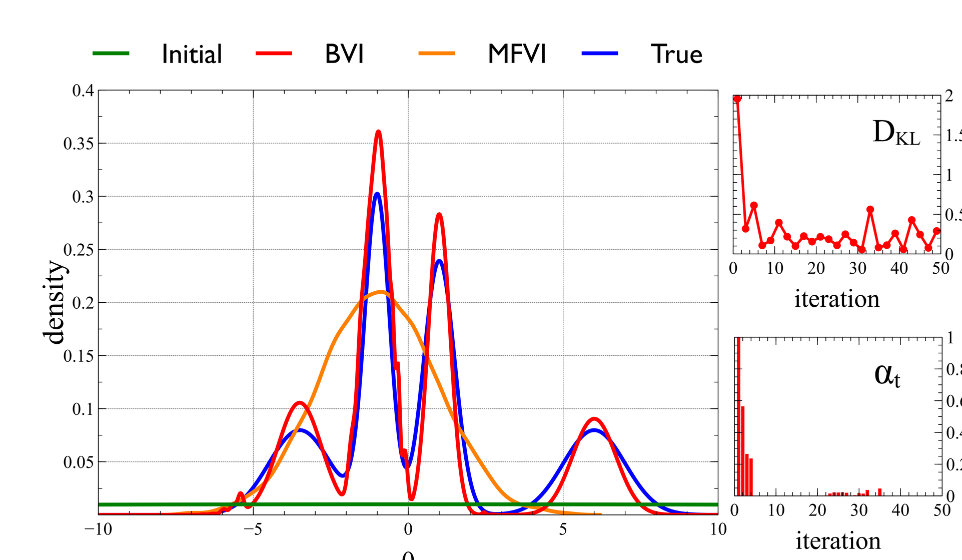


Figure 2: True: Mixture of univariate Gaussians

Logistic Regression Experiment

We run Bayesian logistic regression on the `Nodal` dataset, consisting of $N = 53$ observations of $d = 6$ predictors $\boldsymbol{x}_i$ and a binary response $y_i \in \{-1, +1\}$. We compare to MCMC (as truth) and mean-field VB.

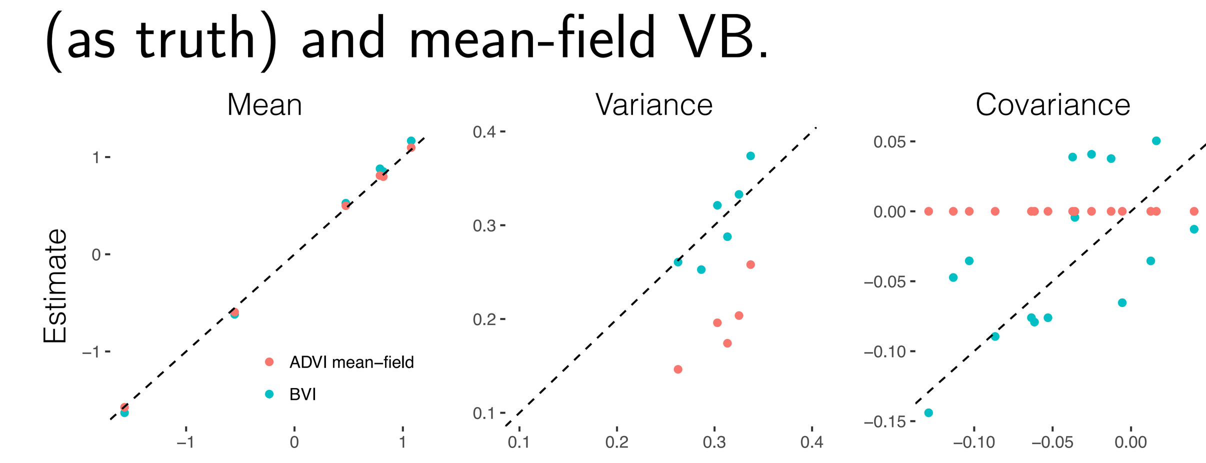


Figure 3: Bayesian logistic regression

References

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Zhang, Tong (2003). "Sequential greedy approximation for certain convex optimization problems". In: *IEEE Transactions on Information Theory* 49.3, pp. 682–691.