Variational inference - reconciling statistical and convergence guarantees

Debdeep Pati
Department of Statistics, University of Wisconsin, Madison

Supported by NSF DMS, NSF CDSE-MSS, NIH R01/R03

July 22, 2025
New Directions in Algebraic Statistics
IMSI, University of Chicago

Bayesian Framework

- Observations $Y^n = (Y_1, \ldots, Y_n)$
- Hidden variables $W^n = (\theta, Z^n)$
  - θ collects all parameters in the model
  - $Z^n = (Z_1, \ldots, Z_n)$ collects all latent variables
- Statistical model:
  - Observed-data likelihood function: $p(Y^n \mid Z^n, \theta)$
  - Latent variable distribution: $p(Z^n \mid \theta)$
  - Prior distribution on parameters: $\pi(\theta)$
- Conduct inference via the joint posterior distribution

$$P\big[d\theta, Z^n \,\big|\, Y^n\big] = \frac{p(Y^n \mid Z^n, \theta)\, p(Z^n \mid \theta)\, \pi(\theta)}{\int_{\Theta \times \mathcal{Z}^n} p(Y^n \mid Z^n, \theta)\, p(Z^n \mid \theta)\, \pi(d\theta)}$$

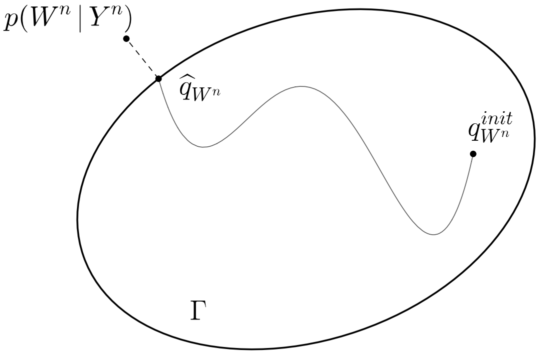
- $\int_{\Theta \times \mathcal{Z}^n} p(Y^n \mid Z^n, \theta)\, p(Z^n \mid \theta)\, \pi(d\theta)$ difficult to obtain beyond simple conjugate settings or low dimensional problems.

How does one compute posterior quantities?

- Markov Chain Monte Carlo (MCMC) sampling avoids computing the denominator
- mixing and scalability issues for “big” data
- Approximate Bayesian inference: Laplace approximation, expectation propagation and variational inference

Variational inference

Feynman (1972), David Mackay (1992, 1995), Hinton and van Camp (1993)



- Let Γ denote a pre-specified family of distributions on $[\Theta, \text{supp}(Z^n)]$
- Idea: approximate the posterior $p(W^n \mid Y^n)$ by a closest member of this family in Kullback-Leibler (KL) divergence

$$\widehat{q}_{W^n} := \underset{q_{W^n} \in \Gamma}{\text{argmin}} \; D_{\text{KL}}\big[q_{W^n}(\cdot) \,\big|\big|\, p(\cdot \mid Y^n)\big]$$

Another perspective: ELBO decomposition

$$\log p(Y^n) = \underbrace{\int_{\mathcal{W}^n} q_{W^n}(w^n) \, \log \frac{q_{W^n}(w^n)}{p(w^n \mid Y^n)} \, dw^n}_{KL[\, q_{W^n}(\cdot) \,||\, p(\cdot \mid Y^n) \,]} + \underbrace{\int_{\mathcal{W}^n} q_{W^n}(w^n) \, \log \frac{p(Y^n \mid w^n) \, p_{W^n}(w^n)}{q_{W^n}(w^n)} \, dw^n}_{L(q_{W^n})}$$

$$\geq L(q_{W^n})$$

- $L(q_{W^n})$ is called the evidence lower bound (ELBO), since it provides a lower bound to the log evidence $\log p(Y^n)$
- $D_{\mathrm{KL}}\left[q_{W^n}(\cdot) \,\middle|\middle|\, p(\cdot \mid Y^n)\right]$ describes the Jensen gap
- KL minimization $\equiv$ ELBO maximization
- Avoids needing to evaluate $p(Y^n)$.

Some commonly used variational families

- Mean-field variational family: consider all joint distribution over $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ that factorizes as $q(\theta) = \prod_{j=1}^{d} q_j(\theta_j)$
- Coordinate ascent: With $F(q) := D(q \,\|\, \pi_n)$, each sub-problem $\text{argmin}_{q_j} F(q_j \otimes q_{-j}^{(t)})$ is convex (however, not jointly)
- Explicit form exploiting the tensorization property of KL divergence

$$q_j^{(t+1)} \propto \exp\left(\int_{\mathcal{X}_{-j}} q_{-j}^{(t)} \log \pi_n \right).$$

Other variational families

- Parametric family such as the exponential family

$$q_{\Theta}(\theta;\, \kappa) = h(\theta)\, \exp\left\{\langle \eta(\kappa),\, T(\theta)\rangle - A(\kappa)\right\}$$

- Normalizing flows (Rezende and Mohamed, 2015)
- Blackbox VI (Ranganath et al 2014)
- Implicit VI (Huszár, 2017)
- Variational Auto-Encoders (Kingma and Welling 2013)
- Mixture of Gaussians (e.g. Zobay, 2014), Implemented using variational boosting (Guo et al 2016, Locatello et al 2017, Miller et al 2019, Campbell and Li, 2019)

Questions of interest

- Statistical Accuracy: Is $\widehat{q}_\Theta$ a good *proxy* for the posterior distribution? Does $\widehat{q}_\Theta$ inherit the good frequentist properties of the posterior?

  P., Bhattacharya and Yang, 2017; Yang, P., Bhattacharya 2019; Wang & Blei, 2019a, 2019b; Zhang and Gao, 2020, Alquier and Ridgeway, 2020, Huggins et al, 2019

- Computational guarantee: Does $\widehat{q}_\Theta^{init}$ converge to $\widehat{q}_\Theta$? Known in specific cases, also in the case of (mean-field) Wasserstein gradient flows Zhang and Zhou, 2017; Mukherjee et al 2018; Locatello et al 2017; Plummer, P and Bhattacharya 2020; Garcia-Trillos and Sanz Alonso, 2020, Lambert et al 2024+, Yao and Yang 2024+, Bhattacharya, P and Yang 2025

Theory for mean-field VB: what to expect

Mean-field VB ignores dependence between parameter blocks. Can not expect full posterior approximation.

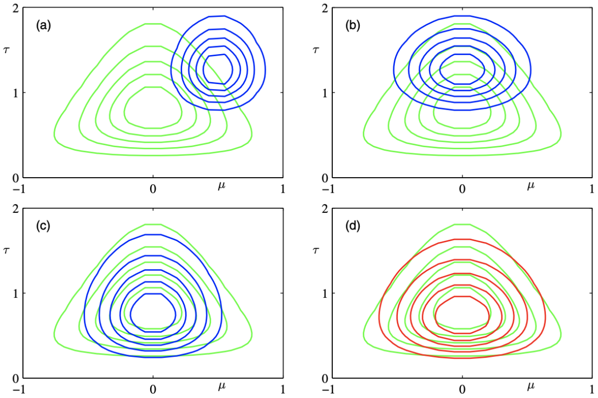


Figure 10.4 Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

Picture credit: Bishop, PMLR

Theory for VB: what to expect

- Statistical Accuracy:
  - The spread of the variational distribution is typically "too small" (e.g. Wang and Titterington, 2005)
  - VB traditionally used for rapidly obtaining point estimates
  - Basic question: Is there any loss of statistical accuracy in terms of convergence rates in using VB?
  - Do point estimates obtained from VB have the same convergence rate as that of the true posterior mean?
  - For non-identifiable models, is ELBO a "good surrogate" for marginal likelihood?
- Computational guarantee (mean-field):
  - Convergence guarantee of a non-convex optimization problem
  - Does initialization play a role?
  - How does the algorithmic convergence rate scale with dimensions?

Example 1: Illustration in sparse regression (Positive result)

Example 1: High-dimensional sparse linear regression

- High-dimensional linear model ($d \gg n$),

$$Y = X\beta + w, \qquad w \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

- Spike and slab mixture prior on β:

$$\pi(\beta_j) = \left(1 - \frac{1}{d}\right)\delta_0 + \frac{1}{d}\mathcal{N}(0, \sigma_\beta^2)$$

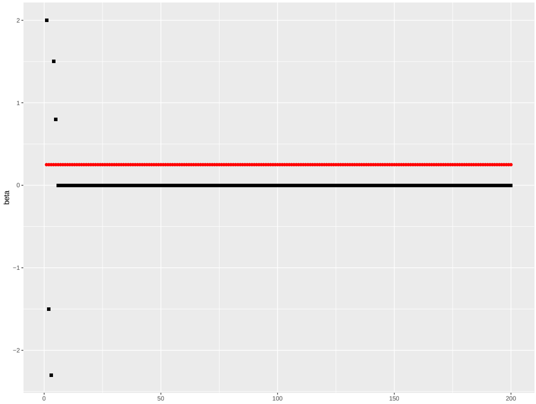
- Mean field variational family to approximate posterior.

$$q(\beta) = \prod_{j=1}^{d} q_{\beta_j}(\beta_j)$$

Fitted regression coefficients ($n = 100$, $d = 200$)

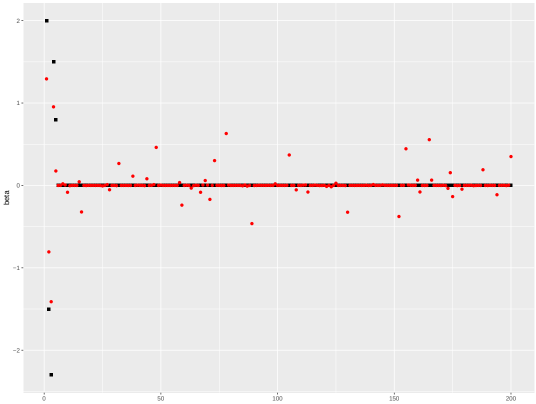
$$Y = X\beta^* + w, \quad w \sim \mathsf{N}(0, \sigma^2)$$

Variational estimate: $\hat{\beta}$ in Red and β^* in Black.



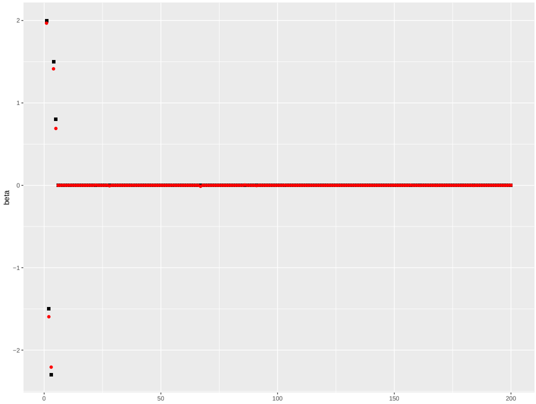
Iteration 1.

Fitted regression coefficients ($n = 100$, $d = 200$)



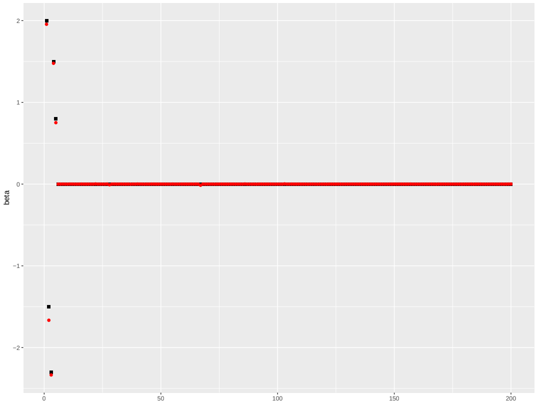
Iteration 2.

Fitted regression coefficients ($n = 100$, $d = 200$)



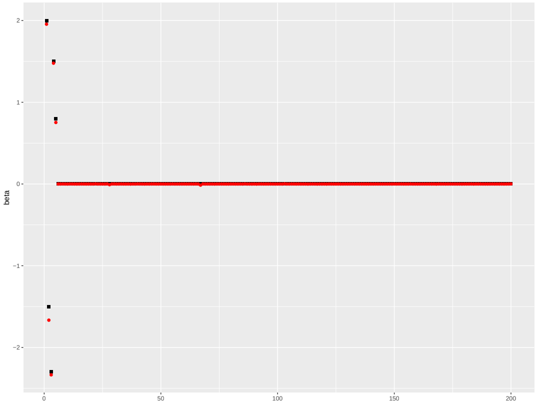
Iteration 3.

Fitted regression coefficients ($n = 100$, $d = 200$)



Iteration 4.

Fitted regression coefficients ($n = 100$, $d = 200$)



Iteration 5.

Example 2: Illustration in linear Gaussian state-space models (Negative result)

Linear Gaussian state space models

- Consider a scalar LGSSM

$$Y_t \mid Z_t \sim \mathcal{N}(bZ_t, \sigma_H^2),\ Z_t \mid Z_{t-1} \sim \mathcal{N}(aZ_{t-1}, \sigma_V^2).$$

- Denote $\theta = (a, b, \sigma_H^2, \sigma_V^2)$ with $a \sim \mathcal{N}(0, \sigma_A^2)$, $b \sim \mathcal{N}(0, \sigma_B^2)$, $\sigma_H^2 \sim \mathsf{IG}(d_{H_1}, d_{H_2})$, $\sigma_V^2 \sim \mathsf{IG}(d_{V_1}, d_{V_2})$.
- Let $W^n = (\theta, Z^n)$ and consider the mean-field family of the form

$$q_{W^n}(W^n) = \left[\prod_{t=1}^{n} q_{Z_t}(Z_t)\right] q_\theta(\theta).$$

Theorem

If the true $a^* \in (0,1)$, then with $\widehat{\theta} = \int \theta \widehat{q}(d\theta)$

$$\lim_{n\to+\infty} \|\widehat{\theta} - \theta^*\| > c$$

for some constant $c > 0$.

Example 3: Illustration in model selection in singular models

Model selection in Bayesian inference

- Observations $Y^n = (Y_1, \ldots, Y_n)$
- k models $\mathcal{M}_j, j = 1, \ldots, k$ where $\mathcal{M}_j := \{\varphi_j(\theta^j), p_j(Y^n \mid \theta^j)\}$
  $\mathcal{M}_j := \{\varphi_j(\theta^j), p_j(Y^{(n)} \mid \theta^j)\}$
- Marginal likelihood or evidence for $\mathcal{M}_j$,
  $m_j(Y^{(n)}) = \int_{\Theta_j} p_j(Y^n \mid \theta^j)\varphi_j(d\theta^j)$ difficult to obtain beyond simple conjugate settings or low dimensional problems.

Laplace approximation

- Marginal likelihood or evidence for $\mathcal{M}_j$, $m_j(Y^n) = \int_{\Omega_j} p_j(Y^n \mid \theta^j)\varphi_j(d\theta^j)$ difficult to obtain beyond simple conjugate settings or low dimensional problems.

- In *regular* parametric models, the Laplace approximation is

$$\log m(Y^n) = \ell_n(\widehat{\theta}_n) - \underbrace{\frac{d \log n}{2}}_{\text{BIC penalty}} + R_n,$$

where $\widehat{\theta}_n$ is the m.l.e. for parameter ξ based on Y^n, d is the parameter dimension, and the remainder term $R_n = O_{P^\star}(1)$.

- Regularity: DGP $f(x)$ and model $p(\cdot \mid \theta)$. If $K(\theta) := D_{\mathrm{KL}}\{f \,\|\, p(\cdot \mid \theta)\}$ has a minimized at a singleton $\theta^\star$ and $-\theta^2/\partial\theta^2 \log p(X \mid \theta)$ is positive definite around θ^*.

Singular models

- The Laplace approximation localizes the integral to a neighborhood of the m.l.e. & applies a 2nd-order Taylor expansion of the log-likelihood to reduce to a Gaussian integral.
- Regularity crucially exploited.
- Singular statistical models: the regularity conditions are not met.

  Mixture models, factor models, hidden Markov models, latent class models, reduced rank regression, neural networks etc. Many of these models routinely appear in economics / econometrics.

Modified approximation for singular models

- In a series of foundational articles, Sumio Watanabe and co-authors (Book: mathematical theory for Bayesian statistics)showed that in singular settings, a more general version of the Laplace approximation is

$$\log m(Y^n) = \ell_n(\theta^\star) - \lambda \log n + (m-1) \log(\log n) + R_n.$$

  assuming the data is generated from $f(y) = p(y \mid \theta^\star)$.

- The quantity $\lambda \in (0, d/2]$ is called the real log-canonical threshold (RLCT) and the integer $1 \leq m \leq d$ its multiplicity.

  When $\lambda = d/2$ and $m = 1 \Rightarrow$ usual Laplace approximation.

- Numerous examples of $\lambda < d/2$ in singular settings (Drton & Plummer, 2017; Watanabe (2009, 2018))

Simple Example

- Singular Model

$$p(y,x \mid a,b,c) = \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[y - \{aS(bx) + cx\}\right]^2\right\} \mathbb{1}_{[-1,1]}(x)$$

$$\varphi(a,b,c) = 1$$

  where $S(x) := x + x^2$ and $(a,b,c) \in [0,1]^3$.
- If the true parameter is $(0,0,0)$,

$$K(a,b,c) = \frac{1}{2}(ab+c)^2 + \frac{1}{6}a^2b^4$$

- For this example $\lambda = 3/4$, $m = 1$.

Mean-field in Original Coordinates (a, b, c)

- Compute the MF approximation to the posterior $q(a, b, c) = q(a)q(b)q(c)$
- For this example the true RLCT and multiplicity are $\lambda = 3/4$, $m = 1$.
- The ELBO recovers

$$\boxed{\text{ELBO}_{MF} \asymp -0.9763 \log(n) + 2.6084}$$

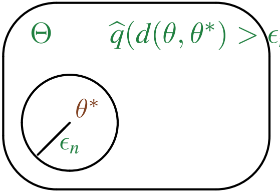
- This is wrong!!!

Back to the drawing board

- Let θ^* denote the (pseudo-)true parameter.
- Variational risk bounds: with high probability under the data-generating distribution, we would want to show

$$\int d^2(\theta, \theta^*)\widehat{q}(d\theta) \leq C\varepsilon_n^2$$

  where d is a distance/divergence measure on the parameter space, and ε_n^2 typically corresponds to the minimax rate (up to a logarithmic term) for the statistical problem.



$$\widehat{q}(d(\theta, \theta^*) > \epsilon_n) \to 0.$$

- If d^2 is convex and $\widehat{\theta} = \int_\Theta \theta \widehat{q}(d\theta)$, then with high prob. $d(\widehat{\theta}, \theta^*) \lesssim \varepsilon_n$.

A simplified setting - no latent variables

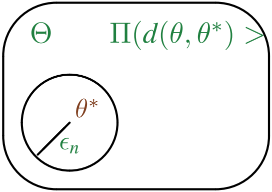
- A key requirement: The posterior itself should be well behaved.

$$\pi_n(\theta) := \frac{\{p(Y^n \mid \theta)\}\, \pi(\theta)}{\int_\Theta \{p(Y^n \mid \theta)\}\, \pi(d\theta)} = \frac{e^{\ell_n(\theta,\theta^*)}\, \pi(\theta)}{\int_\Theta e^{\ell_n(\theta,\theta^*)}\, \pi(d\theta)}$$

where $\ell_n(\theta, \theta^*) = \log\{p(Y^n \mid \theta)/p(Y^n \mid \theta^*)\}$.

- Does the posterior itself concentrate around the (pseudo)-true parameter θ^*?

$$\Pi(d(\theta, \theta^*) > \epsilon_n \mid Y^n) \to 0.$$



- Ghosal and van der Vaart, 2017 lists a few sufficient conditions:
  1. The model should be identifiable in the parameter θ.
  2. The prior should assign enough mass around the θ^*.

First order variational risk bound

- Fix $q \ll \pi$ any probability measure
- Consider

$$D_{\mathrm{KL}}(q, \pi_n) = \underbrace{-\int \ell_n(\theta, \theta^*)\, q(d\theta) + D_{\mathrm{KL}}(q, \pi)}_{\text{-ELBO}} + \log m(Y^n).$$

- Define

$$\boxed{\Psi(q) = \underbrace{-\int \ell_n(\theta, \theta^*)\, q(d\theta)}_{\text{model fit}} + \underbrace{D_{\mathrm{KL}}(q, \pi)}_{\text{penalty}}}$$

Main result

- $h^2(\theta, \theta^*)$ is the squared Hellinger distance between $p(Y^n \mid \theta)$ and $p(Y^n \mid \theta^*)$.

Theorem

Under model identifiability, with high probability,

$$\int_\Theta h^2(\theta, \theta^*)\, \hat{q}(d\theta) \leq C \inf_{q \in \Gamma} \left[\Psi(q)\right] + \textit{Smaller order terms.}$$

- Recall that $\Psi(\cdot)$ is minimized at π_n among all $q \ll \pi$!!
- Minimizing $\Psi(q_\theta)$ within the variational family has the same effect as minimizing the variational Bayes risk

Optimizing the upper bound

- Choose good $q \in \Gamma$ to control $\Psi(q)$

$$q_\delta^{\text{opt}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{B}(\theta^*;\delta)}(\theta)}{\int_\Theta \pi(\theta)\mathbb{I}_{\mathcal{B}(\theta^*;\delta)}(\theta)d\theta}$$

where $\mathcal{B}(\theta^*;\delta) = \{\theta : D_{\mathrm{KL}}(\theta^* \,\|\, \theta) < \delta^2\}$.

- Then

$$\Psi(q_\delta^{\text{opt}}) = \underbrace{-\text{Model fit}}_{\leq n\delta^2} + \underbrace{\text{Penalty}}_{-\log \Pi\{\mathcal{B}(\theta^*;\delta)\}}$$

Theorem

If $-\log \Pi\{\mathcal{B}(\theta^*;\delta)\} \leq h(\delta)$ *and* Γ *is rich enough to contain* q^{opt}*, then* $\Psi(q_\delta^{\text{opt}}) \leq n\delta^2 + h(\delta)$.

Balancing the model fit and the penalty is achieved by choosing δ s.t. $n\delta^2 = h(\delta)$.

Example 1: High-dimensional sparse linear regression

Assumption:

- $\pi_{\beta \mid z^*}$ is continuous assigns sufficient mass around β^*, and the truth β^* is s-sparse.
- Sparse eigen value assumption: For any Cs-sparse vector u, $\|Xu\|^2/\|u\|^2 \geq \mu > 0$.

Theorem

If $s \log d/n \to 0$ as $n \to \infty$, then it holds with probability tending to one as $n \to \infty$ that

$$\left\{ \int h^2\big[p(\cdot \mid \beta) \,\big|\big|\, p(\cdot \mid \beta^*)\big] \, \widehat{q}_\beta(\beta) \, d\beta \right\}^{1/2} \lesssim \sqrt{\frac{s}{n} \log(d\, n)}.$$

Example 2: Structured VB in LGSSM

- Avoid mean-field on Z^n, instead assume

$$q_{W^n}(W^n) = q_{Z^n}(Z^n)\, q_\theta(\theta).$$

- In particular, the computation of univariate and bivariate marginals $q_{Z^n}^{(s)}(Z_k)$ and $q_{Z^n}^{(s)}(Z_k, Z_{k+1})$ at any iteration s can be efficiently carried out using Belief Propagation (BP).

Theorem

If $|a^*| < 1$, then there exists $C, D \geq 0$, such that with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - D/(\log n)$, it holds that

$$\int h^2(\theta, \theta^*)\, \widehat{q}_\theta(d\theta) \leq C \frac{\log n}{n}.$$

Example 3: Structured VB in singular models

- Don't use MF directly on the singular model.
- Use MF after transforming to the "resolved coordinates" ξ Hironaka 1964
- The marginal likelihood approximately becomes

$$\boxed{\int_{[0,1]^d} e^{-n\xi_1^{2k_1}\xi_2^{2k_2}\cdots\xi_d^{2k_d}}\;\xi_1^{h_1}\xi_2^{h_2}\cdots\xi_d^{h_d}\,d\xi}$$

- Using MF on the resolved coordinates, for the non-linear regression example ELBO $= -0.7509\log(n) - 1.5169$.

Mean-Field VI in Dimension $d = 2$

MFVI approximation $\rho(\xi) = \rho_1(\xi_1) \otimes \rho_2(\xi_2)$ to normal form

$$\gamma_K^{(n)}(\xi) \propto \xi_1^{h_1} \xi_2^{h_2} e^{-n\xi_1^{2k_1}\xi_2^{2k_2}}, \quad \xi_1, \xi_2 \in [0, 1].$$

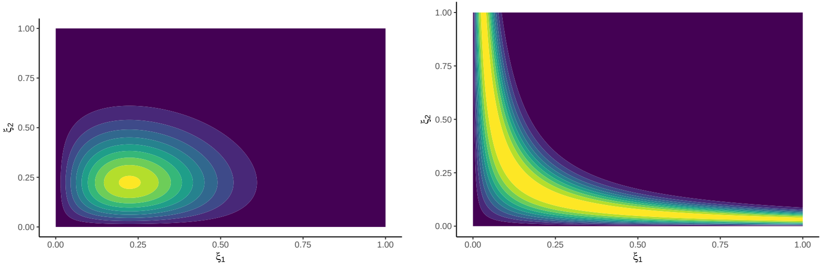
Calculus of variations shows the optimal mean-field approximation is given by marginals,

$$\begin{aligned}
\rho_1^*(\xi_1) \propto \xi_1^{h_1} e^{-n\mu_2^* \xi_1^{2k_1}} \mathbb{1}_{[0,1]}(\xi_1) \quad &= f_{k_1, h_1, n\mu_2^*}(\xi_1), \\
\rho_2^*(\xi_2) \propto \xi_2^{h_2} e^{-n\mu_1^* \xi_2^{2k_2}} \mathbb{1}_{[0,1]}(\xi_2) \quad &= f_{k_2, h_2, n\mu_1^*}(\xi_2),
\end{aligned}$$

where

$$\begin{aligned}
&f_{k,h,\beta}(u) = u^h \exp(-\beta u^{2k}) 1_{[0,1]}(u) / B(k, h, \beta), \\
&B(k, h, \beta) = \int_0^1 x^h \exp(-\beta x^{2k}) dx, \quad G(\lambda, \beta) = \int_0^1 u^{2k} f_{k,h,\beta}(u) du, \\
&\mu_1^* = G(\lambda, n\mu_2^*), \quad \mu_2^* = G(\lambda, n\mu_1^*).
\end{aligned}$$

Mean-Field VI in Dimension $d = 2$



Plot of the optimal mean-field approximation $f_{k_1,h_1,n\mu_2^*}(\xi_1) \otimes f_{k_2,h_2,n\mu_1^*}(\xi_1)$ (LEFT) vs the normal form $\xi_1\xi_2 e^{-n\xi_1^2\xi_2^2}$ (RIGHT) for $h = (1,1), k = (1,1)$.

Example 3: Structured VB in singular models

- Use MF on the resolved coordinates $q(\xi) = q_1(\xi_1) \cdots q_d(\xi_d)$

Theorem

C_1 and C_2 independent of n such that
$-\lambda \log n - C_1 \leq \mathrm{ELBO} \leq -\lambda \log n - C_2$.

- The $\log \log n$ term is missed due to the mean-field approximation, but still surprising since the target after resolving the coordinates, there is still a high dependence structure.

Learn unknown transformation

- Assume $\Gamma : \xi \to \Theta$ be a blowup associated with the resolution of singularities and Q is the mean-field probability on ξ. Let $\tilde{Q} = Q \circ \Gamma^{-1}$, and consider the variational family

$$\mathcal{F}_{tMF} = \{\tilde{q} : \tilde{Q} = Q \circ \Gamma^{-1}, \Gamma \in \mathcal{F}_\Gamma, Q \text{ is MF}\}$$

where $\mathcal{F}_\Gamma$ is a smooth function class.

- Learn Γ using normalizing flows.

Theorem

When optimized over $\mathcal{F}_{tMF}$, $\mathrm{ELBO} = -\lambda \log n + O(1)$.

Recipe for Statistical Accuracy

- Construct likelihood and prior such that posterior should have optimal risk (typically prior should have adequate concentration around the true parameter).
- Variational family should contain (or can approximate) densities of the form

$$q^{\text{opt}}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\mathcal{B}(\theta^*;\delta)}(\theta)}{\int_\Theta \pi(\theta)\mathbb{I}_{\mathcal{B}(\theta^*;\delta)}(\theta)d\theta}.$$

- What this means for mean-field approximation? $\mathcal{B}(\theta^*;\delta)$ should contain a rectangular interval $\mathcal{N}(\theta^*;\delta)$ which has the same prior concentration order.

Algorithmic convergence of coordinate ascent in mean-field

CAVI algorithm

- Suppose $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ with $\mathcal{X}_j \subseteq \mathbb{R}^{m_j}$ and $\sum_{j=1}^{d} m_j = m$. Let

$$\mathcal{Q}_{\text{MF}} := \{q = q_1 \otimes \ldots \otimes q_d \; : \; q \ll \pi_n \text{ and } D_{\text{KL}}(q \,||\, \pi_n) < \infty\}$$

  with q_j a density on $\mathcal{X}_j$ for each $j \in [d]$.
- Denote $F(q) := D_{\text{KL}}(q \,||\, \pi_n)$ to be the variational objective function.
- Write $F(q)$ equivalently as $F(q_j \otimes q_{-j})$ to emphasize dependence on j th coordinate.

CAVI algorithm

- Each sub-problem $\operatorname{argmin}_{q_j} F(q_j \otimes q_{-j}^{(t)})$ is convex (however, not jointly)
- Explicit form exploiting the tensorization property of KL divergence

$$q_j^{(t+1)} \propto \exp\left(\int_{\mathcal{X}_{-j}} q_{-j}^{(t)} \log \pi_n \right).$$

- Iterates analytically tractable in exponential family models
- General framework for convergence?

Ghorbani, Javadi, Montanari 2018; Mukherjee et al. 2018; Huggins et al. 2019; Zhang & Zhou 2020; Alquier & Ridgway 2020; Ghosh, Bhattacharya, P 2022...

Parallel vs. Sequential (2 block case)

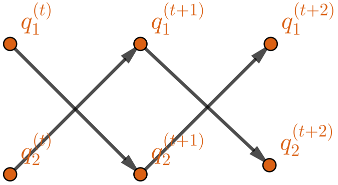


Figure: *Parallel dynamics in 2d with* $q_1^{(t+1)} = \text{argmin}_{q_1} F(q_1 \otimes q_2^{(t)})$ *and* $q_2^{(t+1)} = \text{argmin}_{q_2} F(q_1^{(t)} \otimes q_2)$.

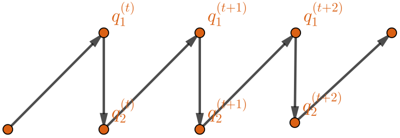


Figure: *Sequential dynamics in 2d with* q_1 *updated first, i.e.,* $q_1^{(t+1)} = \text{argmin}_{q_1} F(q_1 \otimes q_2^{(t)})$ *and* $q_2^{(t+1)} = \text{argmin}_{q_2} F(q_1^{(t+1)} \otimes q_2)$.

Example

- High-dimensional sparse regression
- Coordinate ascent variational inference - try sequential and parallel implementation

```
n <- 100
p <- 100
q <- 20
X=matrix(rnorm(n*p),n,p)
sigmasq=1
E <- rnorm(n,0,sigmasq)
beta=c(rep(1,q),rep(0,p-q))
snr=sd(X%*%beta)/sd(E)
y=X%*%as.matrix(beta) + sigmasq*E
vb_out <- variational_seq(y,X)
vb_out <- variational_par(y,X)
```

Sequential update - estimates

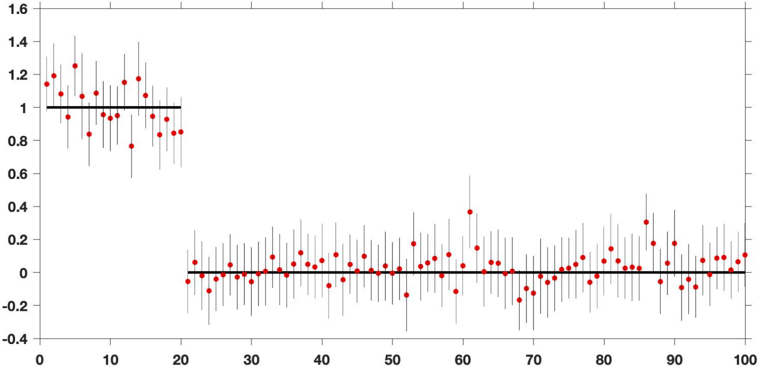


Figure: *Variational mean*, *Pointwise intervals*

Sequential update -tracking ELBO

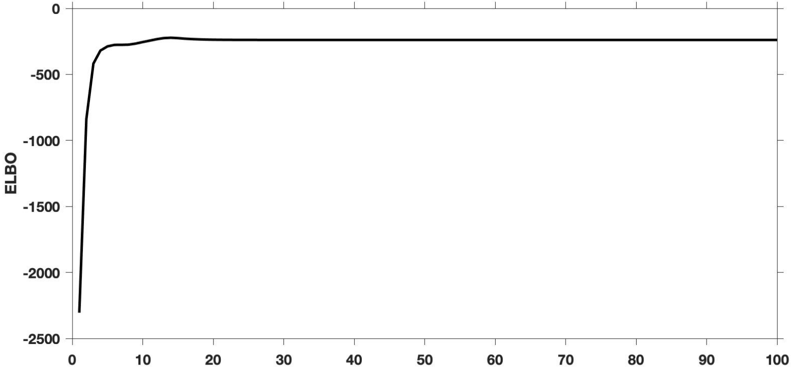


Figure: *ELBO stabilizing after 20 iterations*

Back to the drawing board

- In conditionally conjugate models, the CAVI iterates typically lie inside parametric families $q_j(\cdot \mid \psi_j)$ with $\psi_j \in \Psi_j$ for $j = 1, 2$
- Parallel iterates can be expressed as a finite-dimensional dynamical system $\psi^{(t)} = G(\psi^{(t-1)})$ where $\psi = (\psi_1, \psi_2)$, and $G : \Psi_1 \times \Psi_2 \to \Psi_1 \times \Psi_2$. Similarly for sequential updates.
- One approach: directly analyze this dynamical system. Case specific?

Two-block CAVI

- A key quantity in our theory which captures the interaction between the two blocks:

$$\Delta_n(q_1, q_2) := \int (q_1 - q_1^\star) \otimes (q_2 - q_2^\star) \log \pi_n,$$

where $(q_1^\star, q_2^\star)$ is a global optima of the variational objective F.

- Decomposing $\log \pi_n(\theta_1, \theta_2) = C + V_{n,1}(\theta_1) + V_{n,2}(\theta_2) + V_{n,12}(\theta_1, \theta_2)$,

$$\Delta_n(q_1, q_2) = -\int V_{n,12}(\theta_1, \theta_2)[q_1(\theta_1) - q_1^\star(\theta_1)]\,[q_2(\theta_2) - q_2^\star(\theta_2)]d\theta_1 d\theta_2.$$

- In particular, Δ_n free of the normalizing constant of the target density.

Two-block CAVI

- Let $D_{\mathrm{KL,sym}}(p, q) = D_{\mathrm{KL}}(p \,||\, q) + D_{\mathrm{KL}}(q \,||\, p)$ denote symmetrized KL for densities p, q.

- Define $\mathsf{GCorr}(\pi_n)$ below as the generalized correlation within π_n with respect to the decomposition $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ over families $\mathcal{Q}_1$ and $\mathcal{Q}_2$:

$$\boxed{\mathsf{GCorr}(\pi_n) := \sup_{q_j \neq q_j^\star \in \mathcal{Q}_j} \frac{|\Delta_n(q_1,\, q_2)|}{\sqrt{D_{\mathrm{KL,sym}}(q_1, q_1^\star)\, D_{\mathrm{KL,sym}}(q_2, q_2^\star)}}}.$$

- We show that if $\mathsf{GCorr}(\pi_n) \in (0, 2)$, then parallel / sequential CAVI globally contracts.

Two-block CAVI: global convergence

Theorem

Suppose the target density π_n *satisfies* $\mathrm{GCorr}(\pi_n) \in (0, 2)$. *Then, for any initialization* $q^{(0)} = q_1^{(0)} \otimes q_2^{(0)} \in \mathcal{Q}$ *of the parallel/sequential* CAVI *algorithm, one has a contraction*

$$D_{\mathrm{KL,sym}}(q^{(t+1)}, q^\star) \leq \kappa_n D_{\mathrm{KL,sym}}(q^{(t)}, q^\star),$$

for any $t \geq 0$, *where the contraction constant* $2\kappa_n = \mathrm{GCorr}^2(\pi_n) \in (0, 2)$.

Iterating, for any $t \geq 1$,

$$D_{\mathrm{KL,sym}}(q^{(t)}, q^\star) \leq \kappa_n^t D_{\mathrm{KL,sym}}(q^{(0)}, q^\star).$$

Example (Gaussian)

Suppose $\pi_n \equiv N_p(\theta_0, (nQ)^{-1})$ where Q is a fixed positive definite matrix. Consider a mean-field decomposition $q(\theta) = q_1(\theta_1)\, q_2(\theta_2)$ where we decompose $\theta = (\theta_1, \theta_2)'$ with $\theta_i \in \mathbb{R}^{p_i}$. Partition $\theta_0 = (\theta_{01}, \theta_{02})'$ and

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

Proposition

For a Gaussian target $\pi_n \equiv N_p(\theta_0, (nQ)^{-1})$ with Q positive definite, and a mean-field decomposition as above,
$\mathrm{GCorr}(\pi_n) = 2\|Q_{11}^{-1/2} Q_{12} Q_{22}^{-1/2}\|_2 < 2$.

Remarks

- As a by-product $q^\star = q_1^\star \otimes q_2^\star$ is the unique global minima of F within $\mathcal{Q}$.
- The explicit forms of the updates can be exploited to bound $\mathrm{GCorr}(\pi_n)$ more conveniently.
- Global convergence may not always hold. Formulate a local version of result.

Local convergence

- In the definition of $\mathrm{GCorr}(\pi_n)$, replace $\mathcal{Q}_j$ by $\mathcal{Q}_j^\star(r_0) := \{q_j \in \mathcal{Q}_j : D_{\mathrm{KL}}(q_j^\star \,||\, q_j) \leq r_0\}$ for $r_0 > 0$, and call the resulting quantity $\mathrm{GCorr}(\pi_n; r_0)$.

$$\boxed{\mathsf{GCorr}(\pi_n) := \sup_{q_j \neq q_j^\star \in \mathcal{Q}_j^\star(r_0)} \frac{|\Delta_n(q_1,\, q_2)|}{\sqrt{D_{\mathrm{KL,sym}}(q_1, q_1^\star)\, D_{\mathrm{KL,sym}}(q_2, q_2^\star)}}}\,.$$

Two-block CAVI: local convergence (both parallel and sequential)

Theorem (Two-block CAVI: local contraction)

Suppose there exists $r_0 > 0$ *such that* $\mathrm{GCorr}(\pi_n; r_0) \in (0, 2)$. *Assume that the initialization satisfies* $D_{\mathrm{KL,sym}}(q_j^{(0)}, q_j^\star) \leq r_0$ *for* $j = 1, 2$. *Then, for any* $t \geq 0$,

$$D_{\mathrm{KL,sym}}(q^{(t+1)}, q^\star) \leq \kappa_n D_{\mathrm{KL,sym}}(q^{(t)}, q^\star),$$

with $\kappa_n := \mathrm{GCorr}^2(\pi_n; r_0) \in (0, 2)$.

Examples

Parallel and sequential CAVI concordances for $d = 2$

- Global conditions on $\mathrm{GCorr}(\pi_n)$
  - Multivariate Gaussian target
  - Probit regression
- Characterize r_0 precisely in the condition for $\mathrm{GCorr}(\pi_n; r_0)$
  - Multivariate mean precision
  - Gaussian mixture
  - General expo-family LVM
  - Ising Models - region of convergence corresponds to the Dobrushin regime.

Extension to general d

- Define the generalized correlation between q_j and q_{-j}.

$$\mathrm{GCorr}^{(j)}(\pi_n) = \sup_{q_j \in \mathcal{Q}_j \setminus \{q_j^\star\}} \frac{|\Delta_{j,n}(q_j,\, q_{-j})|}{\sqrt{D_{\mathrm{KL,sym}}(q_j \,||\, q_j^\star)}\,\sqrt{D_{\mathrm{KL,sym}}(q_{-j} \,||\, q_{-j}^\star)}},$$

-

$$\boxed{\mathrm{GCorr}_d(\pi_n) = \max_{j \in [d]} \mathrm{GCorr}^{(j)}(\pi_n)}.$$

- Extension to only parallel case in general d valid with $\mathrm{GCorr}_d(\pi_n) < 2/\sqrt{d-1}$.

General d and parallel and sequential discordances

- Is the condition $\mathrm{GCorr}(\pi_n) < 2/\sqrt{d-1}$ necessary?
- $\pi_n \equiv N_d(0, Q^{-1})$, $Q = (1-\rho)\mathrm{I}_d + \rho 1_d 1_d'$.
- To ensure positive definiteness of Q, assume $-(d-1)^{-1} < \rho < 1$.
- Consider a mean-field approximation $q(\theta) = \prod_{j=1}^d q_j(\theta_j)$.
- $\boxed{\mathrm{GCorr}(\pi_n) < 2/\sqrt{d-1} \Leftrightarrow |\rho| < 1/(d-1)}$.
- The parallel update proceeds as

$$q_j^{(t+1)}(\theta_j) = \mathcal{N}(\theta_j; m_j^{(t+1)}, 1),\ m_j^{(t+1)} = -\rho \sum_{k \neq j} m_k^{(t)},\ j \in [d].$$

- The dynamical system $m^{(t+1)} = \rho(\mathrm{I}_d - 1_d 1_d')m^{(t)}$ converges for $|\rho| < 1/(d-1)$, so our theory is sharp.

Mitigating strategy and concluding remarks

- The parallel scheme itself fails to converge if $1/(d-1) < \rho < 1$.
- We can prove for this particular example that sequential converges within this range.
- For the high-dimensional regression example, stronger conditions on the design are needed for the parallel version for convergence

$$\max_k \sum_{j \in \{S_0\} \setminus \{k\}} \frac{\langle X_j, X_k \rangle^2}{\|X_j\|^2 \|X_k\|^2} \leq 1/(s_0 - 1)$$

where s_0 is the true sparsity and S_0 is the true index set. This forces $s_0^2 = o(n)$ for iid Gaussian.

Merging statistical and computational guarantees

Theorem

Suppose A_n *is such that* $\mathbb{P}_{\theta_0}^{(n)}(A_n) \geq 1 - \delta_n$ *and on* A_n*,*
$\int_\Theta h^2(\theta, \theta_0) q^\star(d\theta) \lesssim \varepsilon_n^2$ *and* $\mathrm{GCorr}(\pi_n; r_0) \in (0, 2)$ *for some* $r_0 > 0$*.*
Assume that the CAVI initialization satisfies $D_{\mathrm{KL},1/2}(q_j^{(0)} \,||\, q_j^\star) \leq r_0/2$*.*
Then, on A_n*, one has*

$$\int_\Theta h^2(\theta, \theta_0) q^{(t)}(d\theta) \lesssim \varepsilon_n^2 \text{ whenever } t \geq t_n := C \log(1/\varepsilon_n)/\log(1/\kappa_n),$$

Concluding remarks and open problems

Recommendations:

- Convergence in latent variable models depend on initialization.
- Avoid parallel CAVI - tends to have cyclical behavior and smaller radius of convergence.

Open problems:

- Convergence Guarantees beyond mean-field?

Co-authors

- Sean Plummer (UArk)
- Honggang Wang (TAMU)
- Anirban Bhattacharya (TAMU)
- Yun Yang (UMD)

References

- Pati, Debdeep, Anirban Bhattacharya, and Yun Yang. “On statistical optimality of variational Bayes.”AISTATS, 2018.
- Yang, Yun, Debdeep Pati, and Anirban Bhattacharya. “α-variational inference with statistical guarantees." Annals of Statistics 48.2 (2020): 886-905.
- Wang, Honggang, Pati, Debdeep, Anirban Bhattacharya, and Yun Yang. “Structured variational inference in Bayesian state-space models ”AISTATS, 2022.
- Bhattacharya, Anirban, Debdeep Pati, and Yun Yang. "On the Convergence of Coordinate Ascent Variational Inference." Annals of Statistics, 2025.
- Ghosh, Indrajit, Anirban Bhattacharya, and Debdeep Pati. “Statistical optimality and stability of tangent transform algorithms in logit models.” JMLR, 2022
- Bhattacharya, Anirban, Debdeep Pati, and Sean Plummer. “Evidence bounds in singular models: probabilistic and variational perspectives.” Statistical Science, 2025.
- Guha, Biraj Subhra, Anirban Bhattacharya, and Debdeep Pati.“Statistical Guarantees and Algorithmic Convergence Issues of Variational Boosting.” IEEE-ICTAI, 2021

Thank You !

Explanation slides

Final target

Theorem

Suppose A_n *is such that* $\mathbb{P}_{\theta_0}^{(n)}(A_n) \geq 1 - \delta_n$ *and on* A_n*,* $\int_\Theta h^2(\theta, \theta_0) q^\star(d\theta) \lesssim \varepsilon_n^2$ *and* $\mathrm{GCorr}(\pi_n; r_0) \in (0, 2)$ *for some* $r_0 > 0$*.* *Assume that the CAVI initialization satisfies* $D_{\mathrm{KL},1/2}(q_j^{(0)} \,||\, q_j^\star) \leq r_0/2$*.* *Then, on* A_n*, one has*

$$\int_\Theta h^2(\theta, \theta_0) q^{(t)}(d\theta) \lesssim \varepsilon_n^2 \text{ whenever } t \geq t_n := C \log(1/\varepsilon_n)/\log(1/\kappa_n),$$

Ising Model on two nodes

- Construct π_n

$$\begin{bmatrix} (0,0) & (0,1) & (1,0) & (1,1) \\ (1-p)/2 & p/2 & p/2 & (1-p)/2 \end{bmatrix},$$

  where $p \in (0,1)$.
- Marginals are $\text{Bernoulli}(0.5)$ each
- If $|\text{logit}(p)| < 2$, CAVI system is globally convergent at $q_1^\star = q_2^\star = \text{Bernoulli}(0.5)$. Indeed, the target density here can be viewed as an Ising model on two nodes, and the condition $|\text{logit}(p)| < 2$ coincides with the Dobrushin regime.
- $|\text{logit}(p)| > 2$, statistically uninteresting since MF minima not at $\text{Bernoulli}(0.5)$ periodic behavior of parallel CAVI

Two-block sequential CAVI

Theorem (Two-block sequential CAVI: local contraction)

Consider $(q_1^{(t)}, q_2^{(t)}) \mapsto (q_1^{(t+1)}, q_2^{(t)}) \mapsto (q_1^{(t+1)}, q_2^{(t+1)})$*, where* q_1 *is updated first.*

Suppose there exists $r_0 > 0$ *such that* $\mathrm{GCorr}(\pi_n; r_0) \in (0,1)$*. Assume that the initialization for* q_1 *satisfies* $D_{\mathrm{KL,sym}}(q_1^{(0)}, q_1^\star) \leq r_0$*, and prepare* $q_2^{(0)} := \mathrm{argmin}_{q_2} F(q_1^{(0)} \otimes q_2)$*. Then, for any* $t \geq 0$*,*

$$D_{\mathrm{KL,sym}}(q_1^{(t+1)} \,||\, q_1^\star) \leq \kappa_n D_{\mathrm{KL,sym}}(q_2^{(t)} \,||\, q_2^\star),$$
$$D_{\mathrm{KL,sym}}(q_2^{(t+1)} \,||\, q_2^\star) \leq \kappa_n D_{\mathrm{KL,sym}}(q_1^{(t+1)} \,||\, q_1^\star),$$

with $\kappa_n := \mathrm{GCorr}^2(\pi_n; r_0) \in (0,1)$*.*

- If the two equations above are satisfied with κ_{1n} and κ_{2n} respectively, then only need $\kappa_{1n}\kappa_{2n} < 1$ for an overall contraction.
- Useful feature in latent variable models.

Example (2-block Gaussian)

Suppose $\pi_n \equiv N_p(\theta_0, (nQ)^{-1})$ where Q is a fixed positive definite matrix. Consider $q(\theta) = q_1(\theta_1)\, q_2(\theta_2)$ with $\theta = (\theta_1, \theta_2)'$ and $\theta_i \in \mathbb{R}^{p_i}$. Partition $\theta_0 = (\theta_{01}, \theta_{02})'$ and

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

The parallel CAVI updates are

$$q_1^{(t+1)}(\theta_1) = \mathcal{N}(\theta_1; m_1^{(t+1)}, (nQ_{11})^{-1}), \quad q_2^{(t+1)}(\theta_2) = \mathcal{N}(\theta_2; m_2^{(t+1)}, (nQ_{22})^{-1}),$$
$$m_1^{(t+1)} = \theta_{01} - Q_{11}^{-1} Q_{12} \big(E_{q_2^{(t)}}(\theta_2) - \theta_{02}\big),\; m_2^{(t+1)} = \theta_{02} - Q_{22}^{-1} Q_{21} \big(E_{q_1^{(t)}}(\theta_1) - \theta_{01}\big).$$

For $q_j \equiv N(m_j, (nQ_{jj})^{-1}),\ j = 1, 2,$

$$\Delta_n(q_1, q_2) = -n\delta_1' Q_{12} \delta_2, \quad \delta_j = \mathbb{E}_{q_j}[\theta_j] - \mathbb{E}_{q_j^\star}[\theta_j] = m_j - m_j^\star.$$

Example (Probit regression)

Suppose $y_i \mid x_i, \beta \overset{ind.}{\sim}$ Bernoulli$(\Phi(x_i'\beta))$ independently for $i \in [n]$. Assume prior $\beta \sim N(0, \kappa^{-1} I_p)$. Augment latent variables $z = (z_1, \ldots, z_n)$ with $y_i = \mathbb{1}(z_i > 0)$ and $z_i \overset{ind.}{\sim} N(x_i'\beta, 1)$. Consider the mean-field decomposition

$$q(\beta, z) = q_{\boldsymbol{\beta}}(\beta)\, q_{\boldsymbol{z}}(z).$$

Let N_1 and N_0 respectively denote univariate truncated normals with truncation region $(0, \infty)$ and $(-\infty, 0)$. The parallel updates are

$$q_{\boldsymbol{\beta}}^{(t+1)}(\beta) = \mathcal{N}_p(\beta; m^{(t+1)}, \Sigma),$$
$$q_{\boldsymbol{z}}^{(t+1)}(z) = \textstyle\prod_{i=1}^n q_i^{(t+1)}(z_i), \;\; q_i^{(t+1)}(z_i) \equiv N_{y_i}(z_i; x_i' m^{(t)}, 1)$$

where $\Sigma = (X'X + \kappa \mathrm{I}_p)^{-1}$, and $m^{(t+1)} = \Sigma X' E_{q_{\boldsymbol{z}}^{(t)}}(z)$.

For $q_{\boldsymbol{\beta}} \equiv N(m, \Sigma)$ with $m \in \mathbb{R}^d$ and $q_{\boldsymbol{z}} = \otimes_{i=1}^n q_i$ with $q_i \equiv N_{y_i}(\alpha_i, 1)$,

$$\Delta_n(q_{\boldsymbol{\beta}}, q_{\boldsymbol{z}}) = \sum_{i=1}^n \left(x_i' m - x_i' m^\star\right)\left(E_{q_i}(z_i) - E_{q_i^\star}(z_i)\right).$$

Example (Mixture model)

Let $x_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(\mu,1)$ with prior $\mu \sim \mathrm{N}(0,\tau_0^{-1})$. Write $x_i \mid z_i, \mu \overset{ind.}{\sim} N(\mu \mathbb{1}(z_1 = 2), 1)$ and $\mathrm{pr}(z_i = 1) = \mathrm{pr}(z_i = 2) = 1/2$. Letting $z = (z_1, \ldots, z_n)'$, consider a mean-field decomposition $q(\mu, z) = q_{\boldsymbol{\mu}}(\mu)\, q_{\boldsymbol{z}}(z)$.

The updates for z lie in the family $q_{\boldsymbol{z}}(z) = \prod_{i=1}^n q_i(z_i)$, where each q_i is a two-point distribution on $\{1,2\}$ with probabilities $(1 - p_i)$ and p_i respectively. Also, the update for μ is of the form $N(m, \tau^{-1})$. Parallel updates:

$$\begin{aligned}
\mathrm{logit}(p_i^{(t+1)}) &= m^{(t)} x_i - \tfrac{1}{2}\left((m^{(t)})^2 + \tfrac{1}{\tau^{(t)}}\right), \\
(m^{(t+1)}, \tau^{(t+1)}) &= \left(\frac{\sum_{i=1}^n p_i^{(t)} x_i}{\tau_0 + \sum_{i=1}^n p_i^{(t)}}, \tau_0 + \sum_{i=1}^n p_i^{(t)}\right).
\end{aligned}$$

For any such $q_{\boldsymbol{\mu}}$ and $q_{\boldsymbol{z}}$,

$$\Delta_n(q_{\boldsymbol{\mu}}, q_{\boldsymbol{z}}) = \sum_{i=1}^n (p_i - p_i^\star)\left[z_i(m)\,(m - m^\star) + \frac{1}{2}\left(\frac{1}{\tau^\star} - \frac{1}{\tau}\right)\right],$$

where $z_i(m) = x_i - (m + m^\star)/2$.