

Explaining Variational Approximations

Ricardo Cortez & Ben Graf

28 Apr 2021

Outline

- Introduction
- Density Transform Approach
- Tangent Transform Approach
- Frequentist Inference
- Conclusion

Introduction

- Variational Approximations
 - deterministic techniques for making approximate inference for parameters in complex models
 - Much faster
 - Limited in Accuracy
- Star of the show in Computer Science
- Monte Carlo has preference in Statistics
- Greater for Bayesian Inference

Density Transform Approach

- Approximates intractable posterior densities with better known and easier to deal with densities
- Two main types of restrictions for the q density:
 - **Product Density Transforms** (non-parametric)
 - where $q(\theta)$ factorizes into $\prod_{i=1}^M q_i(\theta_i)$, for some partition $\{\theta_1, \dots, \theta_M\}$ of θ
 - **Parametric Density Transforms** (parametric)
 - where q is a member of a parametric family of density functions
- Guided by Kullback-Leibler(K-L) divergence
 - Provides a lower bound which can be maximized in order to minimize the K-L divergence between q and $p(\cdot|y)$

Kullback-Leibler Divergence

- Let q be an arbitrary density function over Θ . Then the log of the marginal likelihood satisfies

$$\log(p(\mathbf{y})) \geq \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}$$

- This arises from

$$\int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} \right\} d\boldsymbol{\theta} \geq 0 \quad \text{for all densities } q,$$

with equality iff $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathbf{y})$ almost everywhere (K-L)

- It follows immediately that $p(\mathbf{y}) \geq \underline{p}(\mathbf{y}; q)$, where

$$\underline{p}(\mathbf{y}; q) \equiv \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}$$

is the lower bound on the marginal likelihood

Product Density Transforms

- Suppose q is subject to the product restriction from the previous slide
- It can be shown that the optimal densities satisfy

$$q_i^*(\theta_i) \propto \exp \{E_{-\theta_i} \log p(\mathbf{y}, \theta)\}, \quad 1 \leq i \leq M$$

where $E_{-\theta_i}$ denotes the expectation of the density with q_i removed

- This leads to the algorithm on the next slide to solve for the q_i^*
- Notes:
 - Can show that convergence to at least local optima *guaranteed*
 - If conjugate priors used, then the q_i^* updates reduce to updating parameters in a density family
 - Common to monitor convergence using $\log \underline{p}(\mathbf{y}; q)$

Product Density Transforms

Algorithm 1 Iterative scheme for obtaining the optimal densities under product density restriction. The updates are based on the solutions given on the previous slide.

Initialize: $q_2^*(\theta_2), \dots, q_M^*(\theta_M)$.

Cycle:

$$q_1^*(\theta_1) \leftarrow \frac{\exp\{E_{-\theta_1} \log p(\mathbf{y}, \theta)\}}{\int \exp\{E_{-\theta_1} \log p(\mathbf{y}, \theta)\} d\theta_1},$$
$$\vdots$$

$$q_M^*(\theta_M) \leftarrow \frac{\exp\{E_{-\theta_M} \log p(\mathbf{y}, \theta)\}}{\int \exp\{E_{-\theta_M} \log p(\mathbf{y}, \theta)\} d\theta_M}$$

until the increase in $\underline{p}(\mathbf{y}; q)$ is negligible.

Connection with Gibbs Sampling

- An alternative expression for the q_i^* is

$$q_i^*(\theta_i) \propto \exp \{E_{-\theta_i} \log p(\theta_i \mid \text{rest})\}$$

$$\text{rest} \equiv \{\mathbf{y}, \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_M\}$$

- The distributions $\theta_i \mid \text{rest}$ are called full conditionals in Markov Chain Monte Carlo
- Gibbs sampling uses repeated draws from these
- In fact, product density transforms and Gibbs are tractable in the same scenarios

Product DT Example 1: Normal Random Sample

- Objective is to approximate Bayesian inference for a random sample from a Normal distribution

$$X_i \mid \mu, \sigma^2 \stackrel{\text{ind.}}{\sim} N(\mu, \sigma^2)$$

with conjugate priors

$$\mu \sim N(\mu_\mu, \sigma_\mu^2) \quad \text{and} \quad \sigma^2 \sim \text{IG}(A, B)$$

- The product density transform approximation of $p(\mu, \sigma^2 \mid \mathbf{x})$ is

$$q(\mu, \sigma^2) = q_\mu(\mu)q_{\sigma^2}(\sigma^2)$$

- The optimal densities take the form

$$q_\mu^*(\mu) \propto \exp \left[E_{\sigma^2} \left\{ \log p(\mu \mid \sigma^2, \mathbf{x}) \right\} \right] \quad \text{and}$$

$$q_{\sigma^2}^*(\sigma^2) \propto \exp \left[E_\mu \left\{ \log p(\sigma^2 \mid \mu, \mathbf{x}) \right\} \right]$$

Product DT Example 1: Normal Random Sample

- The resulting estimates are:

$$q_{\sigma^2}^*(\sigma^2) \text{ is InverseGamma} \left(A + \frac{n}{2}, B + \frac{1}{2} E_{\mu} \|\mathbf{x} - \mu \mathbf{1}_n\|^2 \right)$$

$$q_{\mu}^*(\mu) \text{ is Normal} \left(\frac{n \bar{X} E_{\sigma^2} (1/\sigma^2) + \mu_{\mu} / \sigma_{\mu}^2}{n E_{\sigma^2} (1/\sigma^2) + 1/\sigma_{\mu}^2}, \frac{1}{n E_{\sigma^2} (1/\sigma^2) + 1/\sigma_{\mu}^2} \right)$$

and

$$\begin{aligned} \log \underline{p}(\mathbf{x}; q) = & \frac{1}{2} - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \left(\sigma_{q(\mu)}^2 / \sigma_{\mu}^2 \right) \\ & - \frac{\left(\mu_{q(\mu)} - \mu_{\mu} \right)^2 + \sigma_{q(\mu)}^2}{2 \sigma_{\mu}^2} + A \log(B) \\ & - \left(A + \frac{n}{2} \right) \log \left(B_{q(\sigma^2)} \right) + \log \Gamma \left(A + \frac{n}{2} \right) - \log \Gamma(A) \end{aligned}$$

- This leads to the algorithm on the next slide

Product DT Example 1: Normal Random Sample

Algorithm 2 Iterative scheme for obtaining the parameters in the optimal densities q_{μ}^* and $q_{\sigma^2}^*$ in the Normal random sample example.

Initialize: $B_{q(\sigma^2)} > 0$

Cycle:

$$\begin{aligned}\sigma_{q(\mu)}^2 &\leftarrow \left\{ n \left(A + \frac{n}{2} \right) / B_{q(\sigma^2)} + 1 / \sigma_{\mu}^2 \right\}^{-1}, \\ \mu_{q(\mu)} &\leftarrow \left\{ n \bar{X} \left(A + \frac{n}{2} \right) / B_{q(\sigma^2)} + \mu_{\mu} / \sigma_{\mu}^2 \right\} \sigma_{q(\mu)}^2, \\ B_{q(\sigma^2)} &\leftarrow B + \frac{1}{2} \left(\left\| \mathbf{x} - \mu_{q(\mu)} \mathbf{1}_n \right\|^2 + n \sigma_{q(\mu)}^2 \right)\end{aligned}$$

until the increase in $\underline{p}(\mathbf{x}; q)$ is negligible.

Product DT Example 1: Normal Random Sample

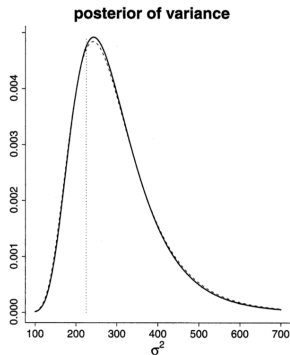
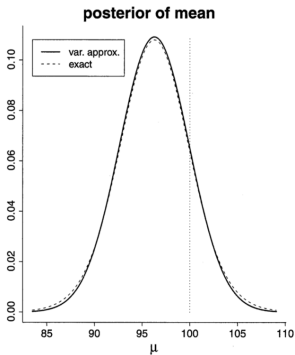
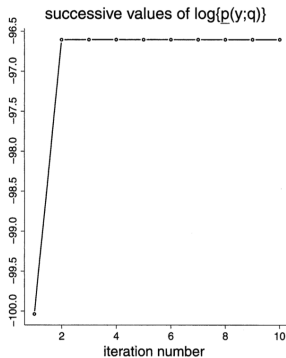
- Upon convergence, the posterior densities are approximated as

$$p(\mu \mid \mathbf{x}) \approx \left\{ 2\pi \left(\sigma_{q(\mu)}^2 \right)^* \right\}^{-1/2} \exp \left[- \left(\mu - \mu_{q(\mu)}^* \right)^2 / \left\{ 2 \left(\sigma_{q(\mu)}^2 \right)^* \right\} \right]$$

$$p\left(\sigma^2 \mid \mathbf{x}\right) \approx \frac{\left(B_{q\left(\sigma^2\right)}^*\right)^{A+\frac{n}{2}}}{\Gamma\left(A+\frac{n}{2}\right)}\left(\sigma^2\right)^{-A-\frac{n}{2}-1} \exp \left(B_{q\left(\sigma^2\right)}^* / \sigma^2\right), \quad \sigma^2>0$$

- Next slide's plots compare product density variational approximations with exact posterior density
 - Sample size $n=20$ from $N(100, 225)$
 - Vague priors chosen: $\mu \sim N(0, 10^8)$, $\sigma^2 \sim IG(\frac{1}{100}, \frac{1}{100})$
 - Initial value $B_{q(\sigma^2)} = 1$
 - Convergence very rapid, accuracy quite good

Product DT Example 1: Normal Random Sample



Product DT Example 2: Linear Mixed Model

- Objective is to approximate Bayesian inference for a random sample from a Gaussian linear mixed model

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}, \mathbf{R} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \quad \mathbf{u}|\mathbf{G} \sim N(\mathbf{0}, \mathbf{G})$$

where

- \mathbf{y} is $n \times 1$ response,
 - $\boldsymbol{\beta}$ is $p \times 1$ fixed effects,
 - \mathbf{u} is random effects,
 - \mathbf{X} and \mathbf{Z} are design matrices, and
 - \mathbf{G} and \mathbf{R} are covariance matrices
- Conjugate priors are

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\beta}^2 \mathbf{I}), \quad \sigma_{u\ell}^2 \sim \text{IG}(A_{u\ell}, B_{u\ell}), 1 \leq \ell \leq r, \quad \sigma_{\varepsilon}^2 \sim \text{IG}(A_{\varepsilon}, B_{\varepsilon})$$

Product DT Example 2: Linear Mixed Model

- The two-component product transform is

$$q\left(\boldsymbol{\beta}, \mathbf{u}, \sigma_{u1}^2, \dots, \sigma_{ur}^2, \sigma_{\varepsilon}^2\right) = q_{\boldsymbol{\beta}, \mathbf{u}}(\boldsymbol{\beta}, \mathbf{u}) q_{\sigma^2}\left(\sigma_{u1}^2, \dots, \sigma_{ur}^2, \sigma_{\varepsilon}^2\right)$$

- This leads to optimal densities

$q_{\boldsymbol{\beta}, \mathbf{u}}^*(\boldsymbol{\beta}, \mathbf{u})$ is a Multivariate Normal density

$q_{\sigma^2}^*$ is a product of $r+1$ Inverse Gamma densities

and...

Product DT Example 2: Linear Mixed Model

$$\begin{aligned} \log p(\mathbf{y}; q) &= \frac{1}{2} (p + \sum_{\ell=1}^r K_{\ell}) - \frac{n}{2} \log(2\pi) - \frac{p}{2} \log(\sigma_{\beta}^2) \\ &\quad + \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(\beta, \mathbf{u})}| - \frac{1}{2\sigma_{\beta}^2} \left\{ \|\boldsymbol{\mu}_{q(\beta)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\beta)}) \right\} \\ &\quad + A_{\varepsilon} \log(B_{\varepsilon}) - (A_{\varepsilon} + \frac{n}{2}) \log(B_{q(\sigma_{\varepsilon}^2)}) \\ &\quad + \log \Gamma(A_{\varepsilon} + \frac{n}{2}) - \log \Gamma(A_{\varepsilon}) \\ &\quad + \sum_{\ell=1}^r \left\{ A_{u\ell} \log(B_{u\ell}) - \left(A_{u\ell} + \frac{K_{\ell}}{2}\right) \log(B_{q(\sigma_{u\ell}^2)}) \right. \\ &\quad \left. + \log \Gamma\left(A_{u\ell} + \frac{K_{\ell}}{2}\right) - \log \Gamma(A_{u\ell}) \right\} \end{aligned}$$

- This leads to the algorithm below and on the next slide

Algorithm 3 Iterative scheme for obtaining the parameters in the optimal densities $q_{\beta, \mathbf{u}}^*$ and $q_{\sigma^2}^*$ in the Bayesian linear mixed model example.

Product DT Example 2: Linear Mixed Model

Initialize: $B_q(\sigma_\varepsilon^2), B_q(\sigma_{u1}^2), \dots, B_q(\sigma_{ur}^2) > 0$

Cycle:

$$\Sigma_{q(\beta, \mathbf{u})} \leftarrow \left\{ \frac{A_\varepsilon + \frac{n}{2}}{B_q(\sigma_\varepsilon^2)} \mathbf{C}^T \mathbf{C} + \text{blockdiag} \left(\sigma_\beta^{-2} \mathbf{I}_p, \frac{A_{u1} + \frac{1}{2} K_1}{B_q(\sigma_{u1}^2)} \mathbf{I}_{K_1}, \dots, \frac{A_{ur} + \frac{1}{2} K_r}{B_q(\sigma_{ur}^2)} \mathbf{I}_{K_r} \right) \right\}^{-1}$$

$$\mu_{q(\beta, \mathbf{u})} \leftarrow \left(\frac{A_\varepsilon + \frac{n}{2}}{B_q(\sigma_\varepsilon^2)} \right) \Sigma_{q(\beta, \mathbf{u})} \mathbf{C}^T \mathbf{y}$$

$$B_q(\sigma_\varepsilon^2) \leftarrow B_\varepsilon + \frac{1}{2} \left\{ \left\| \mathbf{y} - \mathbf{C} \mu_{q(\beta, \mathbf{u})} \right\|^2 + \text{tr} \left(\mathbf{C}^T \mathbf{C} \Sigma_{q(\beta, \mathbf{u})} \right) \right\}$$

$$B_q(\sigma_{u\ell}^2) \leftarrow B_{u\ell} + \frac{1}{2} \left\{ \left\| \mu_{q(\mathbf{u}_\ell)} \right\|^2 + \text{tr} \left(\Sigma_q(\mathbf{u}_\ell) \right) \right\} \quad \text{for } 1 \leq \ell \leq r$$

until the increase in $p(\mathbf{x}; q)$ is negligible.

Product DT Example 2: Linear Mixed Model

- Upon convergence, the posterior densities are approximated as

$p(\beta, \mathbf{u} \mid \mathbf{y}) \approx$ the $N\left(\boldsymbol{\mu}_{q(\beta, \mathbf{u})}^*, \boldsymbol{\Sigma}_{q(\beta, \mathbf{u})}^*\right)$ density function

$p(\sigma_{u1}^2, \dots, \sigma_{ur}^2, \sigma_\varepsilon^2 \mid \mathbf{y}) \approx$
product of the IG $\left(A_{u\ell} + \frac{1}{2}K_\ell, B_{q(\sigma_{u\ell}^2)}^*\right)$, $1 \leq \ell \leq r$, density
functions together with the IG $\left(A_\varepsilon + \frac{1}{2}n, B_{q(\sigma_\varepsilon^2)}^*\right)$ density function

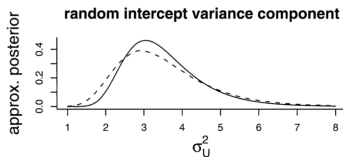
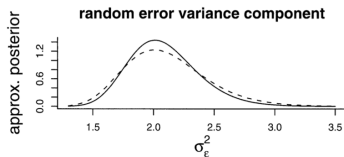
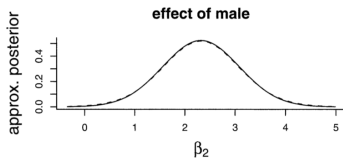
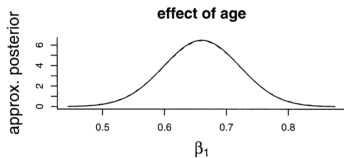
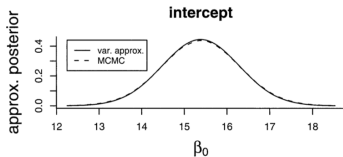
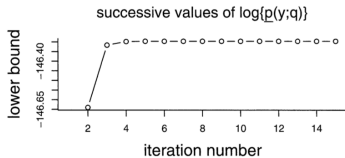
Product DT Example 2: Linear Mixed Model

- Next slide's plots compare product density variational approximations with exact posterior density
 - Data set is longitudinal orthodontic measurements (Pinheiro & Bates, 2000)
 - Random intercept model:

$$\begin{aligned} \text{distance}_{ij} &| U_i \stackrel{\text{ind.}}{\sim} N(\beta_0 + U_i + \beta_1 \text{age}_{ij} + \beta_2 \text{male}_i, \sigma_\varepsilon^2) \\ U_i &| \sigma_u^2 \stackrel{\text{ind.}}{\sim} N(0, \sigma_u^2), \quad 1 \leq i \leq 27, 1 \leq j \leq 4, \\ \beta_i &\stackrel{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \quad \sigma_u^2, \sigma_\varepsilon^2 \stackrel{\text{ind.}}{\sim} \text{IG}(A, B) \end{aligned}$$

- Vague priors chosen: $\sigma_\beta^2 = 10^8$, $A = B = \frac{1}{100}$
- Compared against kernel density estimates using 1M MCMC samples
- Convergence again quite rapid, estimates quite close to MCMC, statistical significance of all parameters

Product DT Example 2: Linear Mixed Model



Parametric DT Example: Poisson Regression

- Now assume q is subject to the parametric restriction
 - Belongs to a specific parametric family that (hopefully) results in a more tractable approximation to the posterior density
- Poisson Regression with Gaussian Transform example
 - Consider the Bayesian Poisson regression model:

$$Y_i | \beta_0 \dots \beta_k \sim \text{Poisson}(\exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}))$$

with priors on the coefficient vector of $\beta \sim N(\mu_\beta, \Sigma_\beta)$

- The marginal likelihood contains an integral that has no closed form solution (intractable):

$$\begin{aligned} p(\mathbf{y}) = & (2\pi)^{-(k+1)/2} |\Sigma_\beta|^{-1/2} \\ & \times \int_{\mathbb{R}^{k+1}} \exp \left\{ \mathbf{y}^T \mathbf{X} \beta - \mathbf{I}_n^T \exp(\mathbf{X} \beta) - \mathbf{I}_n^T \log(\mathbf{y}!) \right. \\ & \left. - \frac{1}{2} (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta) \right\} d\beta \end{aligned}$$

Parametric DT Example: Poisson Regression

- Take $q \sim N(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$

$$\begin{aligned} q(\beta; \boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}) \\ = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_{q(\beta)}|^{-1/2} \exp \left\{ -\frac{1}{2} (\beta - \boldsymbol{\mu}_{q(\beta)})^T \boldsymbol{\Sigma}_{q(\beta)}^{-1} (\beta - \boldsymbol{\mu}_{q(\beta)}) \right\} \end{aligned}$$

- Then the lower bound as defined earlier gives explicitly

$$\begin{aligned} \log \underline{p}(\mathbf{y}; \boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}) \\ = \mathbf{y}^T \mathbf{X} \boldsymbol{\mu}_{q(\beta)} - \mathbf{1}_n^T \exp \left\{ \mathbf{X} \boldsymbol{\mu}_{q(\beta)} + \frac{1}{2} \text{diagonal}(\mathbf{X} \boldsymbol{\Sigma}_{q(\beta)} \mathbf{X}^T) \right\} \\ - \frac{1}{2} (\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta) - \frac{1}{2} \text{tr}(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\Sigma}_{q(\beta)}) \\ + \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(\beta)}| - \frac{1}{2} \log |\boldsymbol{\Sigma}_\beta| + \frac{k+1}{2} - \mathbf{1}_n^T \log(\mathbf{y}!) \end{aligned}$$

Parametric DT Example: Poisson Regression

- From earlier,

$$\log p(\mathbf{y}) \geq \log p(\mathbf{y}; \boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$$

- The optimal variational parameters are found through maximizing this inequality using Newton-Raphson iteration
- This minimizes the K-L divergence and provides the optimal Gaussian density transform q^* as $N(\boldsymbol{\mu}_{q(\beta)}^*, \boldsymbol{\Sigma}_{q(\beta)}^*)$

Tangent Transform Approach

- Not all variational approximations fit into Kullback-Leibler divergence framework
- Tangent transforms work with *tangent-type* representations of concave/convex functions (underpinned by theory of convex duality)

$$\log(x) = \min_{\xi > 0} \{\xi x - \log(\xi) - 1\}, \quad \text{for all } x > 0$$

- The representation implies

$$\log(x) \leq \xi x - \log(\xi) - 1, \quad \text{for all } \xi > 0$$

Tangent Transform Approach

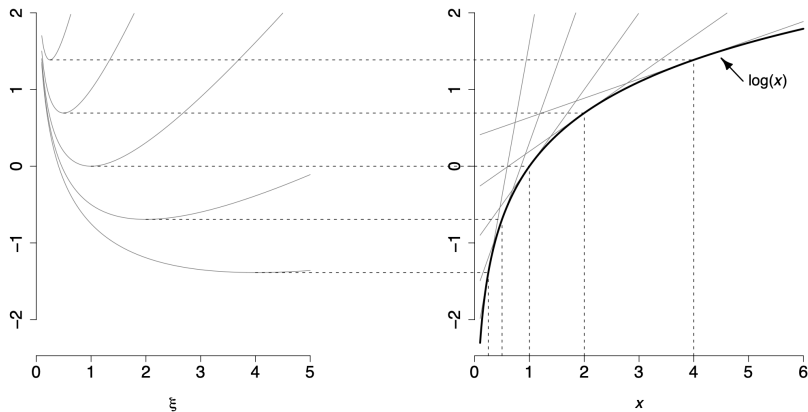


Figure 7: Variational representation of the logarithmic function. Left axes: members of family of functions $f(x, \xi) \equiv \xi x - \log(\xi) - 1$ versus $\xi > 0$, for $x \in \{0.25, 0.5, 1, 2, 4\}$, shown as gray curves. Right axes: For each x , the minimum of $f(x, \xi)$ over ξ corresponds to $\log(x)$. In the x direction the $f(x, \xi)$ are linear and are shown in gray.

TT Example: Bayesian Logistic Regression

- Consider Bayesian logistic regression model

$$Y_i | \beta_0, \dots, \beta_k \stackrel{\text{ind.}}{\sim} \text{Bernoulli} \left([1 + \exp \{ -(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}) \}]^{-1} \right)$$

with priors on the coefficient vector of $\beta \sim N(\mu_\beta, \Sigma_\beta)$

- The posterior density of β is

$$p(\beta | \mathbf{y}) = p(\mathbf{y}, \beta) / \int_{\mathbb{R}^{k+1}} p(\mathbf{y}, \beta) d\beta$$

where

$$p(\mathbf{y} | \beta) = \exp \left[\mathbf{y}^T \mathbf{X} \beta - \mathbf{1}_n^T \log \{ \mathbf{1}_n + \exp(\mathbf{X} \beta) \} \right. \\ \left. - \frac{1}{2} (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta) - \frac{k+1}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_\beta| \right]$$

and the denominator is an intractable integral

TT Example: Bayesian Logistic Regression

- It can be shown $-\log(1 + e^x)$ are the maxima of a family of parabolas:

$$-\log(1 + e^x) = \max_{\xi \in \mathbb{R}} \left\{ A(\xi)x^2 - \frac{1}{2}x + C(\xi) \right\} \quad \text{for all } x \in \mathbb{R}$$

- This is a *tangent-type* representation of a convex function
- From here the derivations proceed similarly to the previous examples

Frequentist Inference

- Frequentist problems that can benefit from variational approximations are rare

Conclusion

- The article's stated goal is to increase statistician's familiarity with variational approximations
- Potential to become major player
 - New methods emerging continually
 - Usefulness increases with problem size, where MCMC becomes untenable
- Does not address accuracy of variational approximations