Advanced Bayesian Learning

Lecture 5 - Mean field and stochastic variational inference

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Overview

- Variational inference (VI)
- Mean-field VI
- Stochastic VI
- Fixed form VI
- Stochastic gradients and variance reduction
- Automatic differentiation

Variational inference

- Literature:
  - *Variational Inference: A Review for Statisticians*, JASA article by Blei et al (2017).
  - *A practical tutorial on Variational Bayes* - notes by Minh-Ngoc Tran at Sydney University.
- Aim: approximate $p(\boldsymbol{\theta}|\boldsymbol{y})$ with a (simpler) distribution $q(\boldsymbol{\theta})$.
- **Laplace approximation** from optimization:

$$q(\boldsymbol{\theta}) = N\left[\tilde{\boldsymbol{\theta}}, \left(-\nabla\nabla \log p(\boldsymbol{\theta}|\boldsymbol{y})|_{\tilde{\boldsymbol{\theta}}}\right)^{-1}\right]$$

- **Kullback-Leibler divergence** of $g(x)$ from $f(x)$

$$\mathrm{KL}(f\,\|g) = \int \ln \frac{f(x)}{g(x)} f(x)dx = \mathbb{E}_f\left(\ln \frac{f(x)}{g(x)}\right)$$

- **Properties** of KL:
  - $\mathrm{KL}(f\,\|g) \geq 0$
  - $\mathrm{KL}(f\,\|g) \neq \mathrm{KL}(g\,\|f)$ in general. First density is the judge.

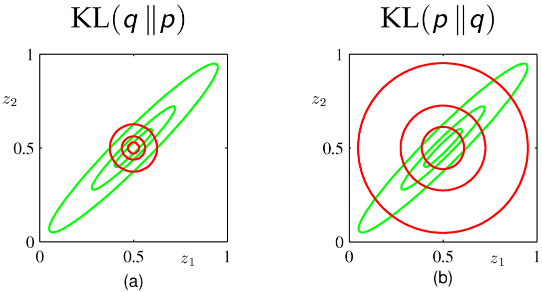
Variational inference

- **VI**: approximate $p(\boldsymbol{\theta}|\boldsymbol{y})$ by $q(\boldsymbol{\theta}) \in \mathcal{Q}$

$$q^{\star}(\theta) = \underset{q(\theta)\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}(q\,\|\,p) = \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\theta$$

- Turns an inference problem, $p(\boldsymbol{\theta}|\boldsymbol{y})$, into **optimization**.
- **Ideal**:
  - let $\mathcal{Q}$ be **large enough to approx** $p(\boldsymbol{\theta}|\boldsymbol{y})$ well
  - let $\mathcal{Q}$ be **small enough for efficient optimization**
- **Early VI**: use restrictive $\mathcal{Q}$ and live with poor approximation.
  - Location of $p(\boldsymbol{\theta}|\boldsymbol{y})$ is fairly correct.
  - Underestimates the variance (badly).
- **Modern VI**: use larger $\mathcal{Q}$ + better optimization algorithms + stochastic gradients.

KL - forward or reverse[1]



Green contours = True Gaussian posterior
Red contours = Circular Gaussian approximation

[1] From Bishop's book *Pattern Recognition and Machine Learning*, Springer.

ELBO - evidence lower bound

- $\mathrm{KL}(q, p)$ is intractable when $p(\boldsymbol{\theta}|\boldsymbol{y})$ is intractable, but

$$\begin{aligned}\mathrm{KL}(q\,\|p) &= \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\theta = \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y})q(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} d\theta \\ &= -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\theta + \int p(\boldsymbol{y})q(\boldsymbol{\theta}) d\theta\end{aligned}$$

- Hence $\mathrm{KL}(q\,\|p) = -\mathrm{LB}(q) + p(\boldsymbol{y})$ where

$$\mathrm{LB}(q) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\theta$$

  is a lower bound for the marginal likelihood $p(\boldsymbol{y})$

$$\mathrm{KL}(q\,\|p) \geq 0 \Longrightarrow \mathrm{LB}(q) \leq p(\boldsymbol{y})$$

- $\mathrm{LB}(q)$ sometimes called evidence lower bound (**ELBO**).

Mean field approximation

- **Mean field VI** is based on factorized approximation:

$$q(\theta) = \prod_{j=1}^{p} q_j(\theta_j)$$

- **No specific functional forms** are assumed for the $q_j(\theta)$.

- **Optimal densities** can be shown to satisfy (MNT Notes):

$$q_j(\theta) \propto \exp\left(E_{-\theta_j} \ln p(\mathbf{y}, \theta)\right)$$

  where $E_{-\theta_j}(\cdot)$ is the expectation with respect to $\prod_{k \neq j} q_k(\theta_k)$.

- **Structured mean field approximation**. Group subset of parameters in tractable blocks. Similar to Gibbs sampling.

Mean field VI - algorithm

- Initialize: $q_2^*(\theta_2), \ldots, q_M^*(\theta_p)$
- Repeat until convergence:
  - $q_1^*(\theta_1) \leftarrow \frac{\exp\left[E_{-\theta_1} \ln p(\mathbf{y},\theta)\right]}{\int \exp\left[E_{-\theta_1} \ln p(\mathbf{y},\theta)\right] d\theta_1}$
  - $q_2^*(\theta_2) \leftarrow \frac{\exp\left[E_{-\theta_2} \ln p(\mathbf{y},\theta)\right]}{\int \exp\left[E_{-\theta_2} \ln p(\mathbf{y},\theta)\right] d\theta_2}$
  - $\vdots$
  - $q_p^*(\theta_p) \leftarrow \frac{\exp\left[E_{-\theta_p} \ln p(\mathbf{y},\theta)\right]}{\int \exp\left[E_{-\theta_p} \ln p(\mathbf{y},\theta)\right] d\theta_p}$
- No assumptions about parametric form of the $q_j(\theta)$.
- Optimal $q_j(\theta)$ often turn out to be known distributions.
- Just update hyperparameters in the optimal densities.

Mean field VI

- Alternative formulation that connects to **Gibbs sampling**

$$q_i^*(\theta_i) \propto \exp\left[E_{-\theta_i} \ln p(\theta_i|\theta_{-i}, \boldsymbol{y})\right]$$

where $p(\theta_i|\theta_{-i}, \boldsymbol{y})$ is the full conditional posterior of θ_i.

- **Structured mean field VI**. Group parameters in tractable blocks.

- Make life easy. When deriving $q_{\theta_1}^*(\theta_1)$:
  - ignore additive terms in $\ln p(\theta_1, \theta_2, \theta_3, \mathbf{y})$ not involving θ_1.
  - mean-field: $\mathbb{E}_{-\theta_1} f(\theta_2) g(\theta_3) = \mathbb{E}_{q_2(\theta_2)} f(\theta_2) \cdot \mathbb{E}_{q_3(\theta_3)} g(\theta_3)$.
  - And of course $\mathbb{E}_{-\theta_1} f(\theta_1) = f(\theta_1)$

Mean field approximation - Normal model

- **Model**: $X_i | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$.
- **Prior**: $\theta \sim N(\mu_0, \tau_0^2)$ *independent* of $\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$.
- **Mean-field approximation**: $q(\theta, \sigma^2) = q_\theta(\theta) \cdot q_{\sigma^2}(\sigma^2)$.
- Optimal densities

$$q_\theta^*(\theta) \propto \exp\left[E_{q(\sigma^2)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$
$$q_{\sigma^2}^*(\sigma^2) \propto \exp\left[E_{q(\theta)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$

Normal model - VB algorithm

- **Variational density for σ^2**

$$\sigma^2 \sim Inv - \chi^2 \left(\tilde{\nu}_n, \tilde{\sigma}_n^2\right)$$

where $\tilde{\nu}_n = \nu_0 + n$ and $\tilde{\sigma}_n = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^{n}(x_i - \tilde{\mu}_n)^2 + n \cdot \tilde{\tau}_n^2}{\nu_0 + n}$

- **Variational density for θ**

$$\theta \sim N\left(\tilde{\mu}_n, \tilde{\tau}_n^2\right)$$

where

$$\tilde{\tau}_n^2 = \frac{1}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

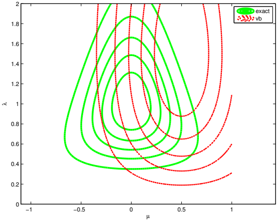
$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1 - \tilde{w})\mu_0,$$

where

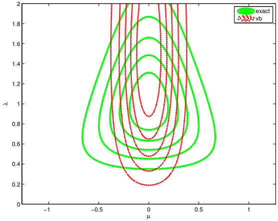
$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

Normal example $(\lambda = 1/\sigma^2)$[2]

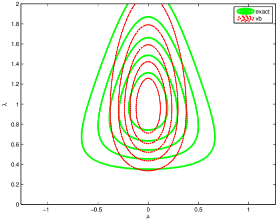
Initial values



After updating q_μ

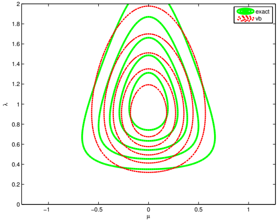


After updating q_{σ^2}



$<$

At convergence



[2] From Bishop's book *Pattern Recognition and Machine Learning*, Springer.

Probit regression[3]

- **Model**:

$$\Pr(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i^T \beta)$$

- **Prior**: $\beta \sim N(0, \Sigma_\beta)$. For example: $\Sigma_\beta = \tau^2 I$.
- **Latent variable formulation** with $u = (u_1, ..., u_n)'$

$$\mathbf{u}|\beta \sim N(\mathbf{X}\beta, 1)$$

and

$$y_i = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$

- Factorized **variational approximation**

$$q(\mathbf{u}, \beta) = q_{\mathbf{u}}(\mathbf{u}) q_\beta(\beta)$$

[3] From Ormerod and Wand (2010). *Explaining Variational Approximation, Amer Stat.*

VI for probit regression

VI posterior

$$\beta \sim N\left(\tilde{\mu}_\beta, \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

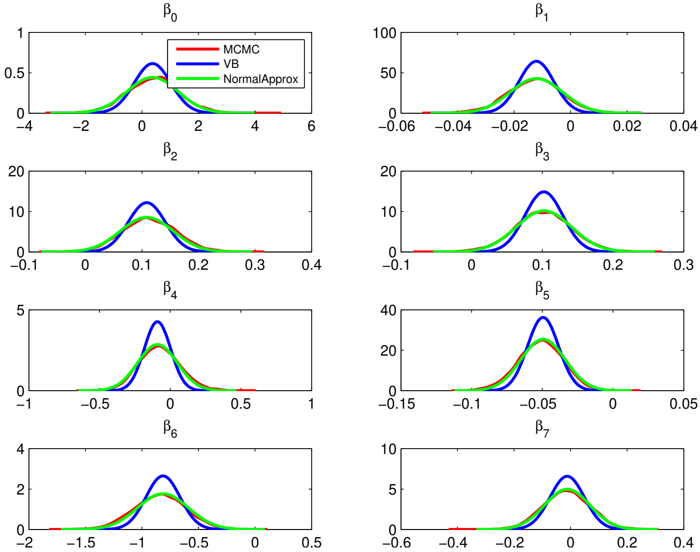
where

$$\tilde{\mu}_\beta = \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\mathbf{X}^T\tilde{\mu}_\mathbf{u}$$

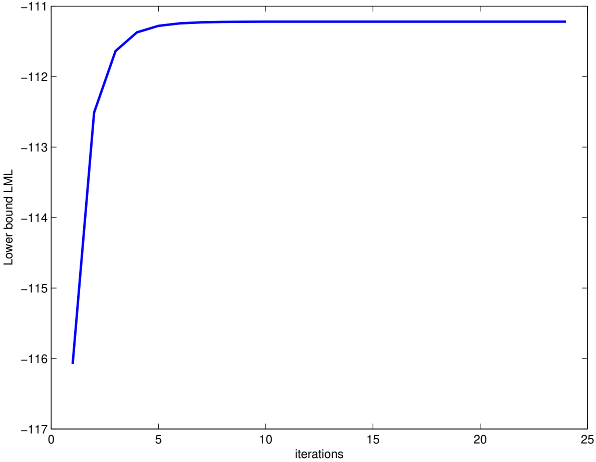
and

$$\tilde{\mu}_\mathbf{u} = \mathbf{X}\tilde{\mu}_\beta + \frac{\phi\left(\mathbf{X}\tilde{\mu}_\beta\right)}{\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right)^{\mathbf{y}}\left[\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right) - \mathbf{1}_n\right]^{\mathbf{1}_n - \mathbf{y}}}.$$

Probit example (n=200 observations)



Probit example



VI and exponential families

- **Exponential family** with sufficient statistics $\boldsymbol{t}(\boldsymbol{x})$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\eta(\boldsymbol{\theta})^T \boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\theta})\right\}$$

- Suppose full conditional posterior is in the exponential family

$$p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{y}) = h(\theta_j) \exp\left\{\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\theta_j - a\left(\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right)\right\}$$

- Mean-field VI update

$$\begin{aligned} q(\theta_j) &\propto \exp\left\{\mathbb{E}_{-j} \log p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right\} \\ &= \exp\left\{\log h(\theta_j) + \mathbb{E}_{-j}\left[\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right]\theta_j - \mathbb{E}_{-j}\left[a\left(\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right)\right]\right\} \\ &\propto h(\theta_j) \exp\left\{\mathbb{E}_{-j}\left[\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right]\theta_j\right\} \end{aligned}$$

- Each $q(\theta_j)$ has same exponential family as its full conditional but with parameter $\mathbb{E}_{-j}\left[\eta_j(\boldsymbol{\theta}_{-j}, \boldsymbol{y})\right]$.

Digression - Conjugate prior for expon family

- **Exponential family** in the canonical parametrization

$$p(x|\theta) = h(x) \exp\left(\theta^T \mathbf{t}(x) - A(\theta)\right)$$

- **Likelihood**

$$p(x_1, ..., x_n|\theta) = \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left(\theta^T \sum_{i=1}^{n} \mathbf{t}(x_i) - nA(\theta)\right)$$

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0) \exp\left(\theta^T \tau_0 - n_0 A(\theta)\right),$$

where τ_0 and n_0 are prior hyperparameters and $H(\tau_0, n_0)$ is the normalizing constant which is known to exist if $n_0 > 0$.

Digression - Posterior in exponential family

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0) \exp\left(\theta^T \tau_0 - n_0 A(\theta)\right)$$

- **Posterior**

$$p(\theta | x_1, \ldots, x_n) \propto \exp\left[\theta^T \left(\tau_0 + \sum_{i=1}^{n} \mathbf{t}(x_i)\right) - (n_0 + n) A(\theta)\right]$$

- **Prior-to-posterior updating**

$$\tau_0 \Longrightarrow \tau_n = \tau_0 + \sum_{i=1}^{n} \mathbf{t}(x_i)$$

$$n_0 \Longrightarrow n_0 + n$$

Digression - Bernoulli as exponential family

- **Exponential family** in the non-canonical parametrization

$$p(x|\theta) = h(x)\exp\left(\phi(\theta)^T \mathbf{t}(x) - A(\theta)\right)$$

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0)\exp\left(\phi(\theta)^T \tau_0 - n_0 A(\theta)\right)$$

- **Bernoulli likelihood**

$$\begin{aligned} p(x_1, \dots, x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \exp\left(\log\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{n} x_i - n\log\left(\frac{1}{1-\theta}\right)\right) \\ &= \exp\left(\phi(\theta)\sum_{i=1}^{n} x_i - nA(\theta)\right) \end{aligned}$$

where $\phi = \log\left(\frac{\theta}{1-\theta}\right)$ and $A(\theta) = \log\left(\frac{1}{1-\theta}\right)$.

- **Conjugate prior $p(\phi)$**

$$\exp\left(\phi(\theta)\tau_0 - n_0 A(\theta)\right) = \exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_0 - n_0\log\left(\frac{1}{1-\theta}\right)\right) = \theta^{\tau_0}(1-\theta)^{n_0-\tau_0}$$

Stochastic variational inference, Blei et al 2017

- **Mixture**: $\Pr(z_i = k) = \omega_k$ and $x_i|(z_i = k) \sim N(x|\mu_k, \sigma_k^2)$.

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega}) = p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega}) \prod_{i=1}^{n} p(x_i|z_i, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(z_i|\boldsymbol{\omega})$$

- **Global parameters**: $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega})^T$.
- **Local parameters**: z_i (**latents**). z_i is local to x_i.
- **Mean field VI** for **local parameter models** iterates:
  - Update the variational factor $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ for global parameters.
  - Update the variational factor $q(z_i|\varphi_i)$ for each local z_i.
- **Stochastic VI** (Blei et al 2017) for large data with latents:
  - Subsample a data point $s \in \{1, ..., n\}$ and update $q(z_s|\varphi_s)$.
  - Update the variational factor $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$ for global parameters.